UNIVERSITY
OF FLORIDA
LIBRARIES

1853

Digitized by the Internet Archive
in 2013

http://archive.org/details/economictheoryor00leib

Economic Theory and Organizational Analysis

Economic Theory and Organizational Analysis

Harvey Leibenstein
UNIVERSITY OF CALIFORNIA,
BERKELEY

HARPER & BROTHERS, PUBLISHERS, NEW YORK

ECONOMIC THEORY AND ORGANIZATIONAL ANALYSIS
Copyright © 1960 by Harvey Leibenstein

Printed in the United States of America
All rights in this book are reserved. No part of the book may be used or reproduced in any manner whatsoever without written permission except in the case of brief quotations embodied in critical articles and reviews. For information address Harper & Brothers, 49 East 33rd Street, New York 16, N.Y.

D-K

Library of Congress catalog card number: 60–7004

Contents

Preface vii

PART I. MICROECONOMICS

1. Some Introductory Concepts 3
2. Demand Theory 13
3. The Competitive Firm in the Short Run 24
4. The Competitive Economy in the Long Run 41
5. Monopoly and Monopolistic Competition 57
6. The Oligopoly Problem 73

PART II. SPECIALIZATION

7. On the Meaning of Specialization 89
8. Some Determinants of the Degree of Specialization 101

PART III. ORGANIZATIONAL ANALYSIS

9. Roles and Role-Taking in Organizations 119
10. Toward the Definition of an Organization 136
11. The Authority-Communication Mechanism 155
12. Decision-Making and Motivation 168
13. The Functioning of the Firm and Organization Equilibrium 194
14. Some Economic Aspects of Organization 212
15. Considerations Bearing on the Determination of an Efficient Organization Structure 234

PART IV. IMPLICATIONS FOR ECONOMIC THEORY

16. The Firm Growth Hypothesis 265
17. Organizational Objectives and the Profit Maximization Hypothesis 274

18. Organization Costs and the Scale of Operations 283
19. Cost Curves, Growth, and the Long-Run Theory of the Firm 295
20. Notes on Organization and Oligopoly Behavior 317

Index of Names 345

Index of Subjects 347

Preface

THE PURPOSE of this book is to consider two types of theoretical approaches that are closely related: microeconomic theory and organizational analysis, with emphasis on the latter. Microeconomic theory is a well-developed subject and only the elements of that theory need be presented here to be able to contrast the two approaches and, where possible, to compare some of their results. Organization theory, on the other hand, is not well developed, and consequently a considerable portion of this book will be devoted to the development of a theory of organization built on the foundation of a few critical ideas.

There was a time when economists attempted to write comprehensive treatises, but the expansion of the field makes such attempts exceedingly difficult and, in an important sense, unnecessary. In order to appreciate a scheme of analysis, one need not know all the possible variations nor all the conceivable applications of a given analytical technique. The technique can usually be understood by a consideration of a limited number of examples.

In this connection we recall Keynes' remarks in his introduction to the *Cambridge Economic Handbooks,* ". . . Economics does not furnish a body of settled conclusions. . . . It is an apparatus of the mind, a technique of thinking, which helps its possessor to draw correct conclusions." If the essence of a theory is an "apparatus of the mind," then clearly comprehensiveness is not essential to understanding the approach. I emphasize this because in this book no attempt will be made to be exhaustive. Far from it; this book is written very much in the spirit of the quotation. Obviously we want to say enough about a theory to expose clearly the underlying system of analysis involved, but it would be otiose to try to say everything that could be said on the subject. Furthermore, once we are relieved of the passion for exhaustiveness, we can be deliberate and rounded in discussing those elements of the problem that we choose for our illustrations.

Part I is a brief review of price, costs, and output theory, while Part II analyzes the phenomenon of specialization central to both cost and output theory and to organizational analysis. It is a bridge between the two fields. In conventional economic theory, behavior is assumed to take place between such basic entities as households and firms. But these entities are not the smallest elements that might be considered. Within most firms there are interactions between individuals which determine company decisions. Furthermore, the decisions imply sets of activities that such individuals must carry out in a more or less organized fashion through and for the basic entity that we call the firm. Thus, if we want to go a little more deeply into the analysis of the phenomena considered by contemporary economic theory, we have to analyze behavior within the basic entities as well as between them. It is from this point of view that the central elements of a theory of organization are presented in Part III.

Let us return to economic theory for a moment. It is probably true that the most helpful aspects of microeconomic theory do not lie in its conclusions, but rather in the fact that the theory provides a small but closely related set of concepts that can be employed with effect in the discussion and analysis of practical problems. Similarly, it may be hoped that the construction of a small set of closely related concepts based on a limited set of pivotal ideas may turn out to be useful in organizational analysis. But be that as it may, this is the aim of the system of organizational analysis presented here.

The basic conceptual scheme developed here is open rather than closed. It is constructed in such a way that it can quite readily be elaborated upon or built onto in order to take care of complexities not considered here. Certainly in the present state of our knowledge about organizational behavior it would seem more appropriate to consider an open rather than a closed system.

To be more specific, we shall look upon the firm as an organization whose problem is, in part, to organize its members in such a way as to best—or at least satisfactorily—carry out the organization's activities. We postulate that the organization must work on the basis of incomplete knowledge and that part of its effort must be directed toward the gathering and processing of information. In addition, we neither assume that the agents of production come in all conceivable sizes, nor that they are infinitely divisible, but that they are, in fact, often of a lumpy nature. Hence the problem of organization is, in part, a problem of organizing these lumpy agents into workable constellations in such a way that their relations vis-à-vis each other are consistent with the performance aimed at by those that direct the organization. Throughout, the emphasis will be on the human agent in the production process.

The elements of the theory of organization to be presented will not be constructed so narrowly as to be applicable only to the theory of the firm. They are sufficiently general so as to be applicable to many types of nonprofit-pursuing organizations. The emphasis will be—if we may be permitted a biological analogy—on the anatomy and physiology of organization, rather than on the pathology, despite the likelihood that many of the most interesting organizational problems fall into the latter category (e.g., Parkinson's Law). But an appreciation of the former should be helpful for an understanding of the latter.

In Part IV some of the broad ideas developed in the other parts of the book are applied to the theory of the firm and the industry in the hope of shedding light on some aspects of firm and industry behavior. More specifically, we shall consider among other things: the meaning and nature of firm objectives, the concept of entrepreneurship, the determination of firm size, the reasons for firm growth, the relation between administrative costs and size, and the theory of oligopoly. It should be emphasized that no attempt has been made to say everything that can possibly be said on all these matters, but only an attempt to show how some of the concepts developed in earlier parts of the book may shed some light on these issues.

In thinking about the problem of organization and the theory of the firm, I was stimulated by, and am indebted to, the writings of Chester Barnard and Herbert A. Simon, and the writings of some of my colleagues, especially Professors J. S. Bain, R. A. Gordon, and A. G. Papandreou. I am also indebted to: Professor William Baumol of Princeton for a number of conversations that we had about oligopolistic firm behavior that were both stimulating and enlightening; Professor Hyman Minsky for reading and criticizing some of the early chapters of the book; Professors William Baumol, R. A. Gordon, and Julius Margolis for taking time out from their considerable academic and research activities to read earlier drafts of the manuscript and for their useful comments and suggestions.

Part of the manuscript was written while I was a Social Science Research Council Faculty Research Fellow, and I am happy to express my gratitude to the S.S.R.C. for the research opportunities that they have made possible. I am also indebted to the Bureau of Business and Economic Research of the University of California at Berkeley for aid in the typing of the manuscript.

There are also psychological aspects to the writing of a book. In this connection I am grateful to my wife, Margaret Libnic Leibenstein, for encouragement when my enthusiasm flagged and for her patience when I was overly zealous.

Although others have contributed toward the merits of this book, it is as usual up to the author to take responsibility for the errors that may be contained therein.

HARVEY LEIBENSTEIN

January, 1960
Berkeley, California

Part I

Microeconomics

1

Some Introductory Concepts

IN ECONOMICS we consider the behavior of certain basic decision-making units such as households and firms, and their relation to the behavior of markets. We are not concerned whether the decision-making units are large or small, whether they are single individuals or entities made up of a large number of individuals, e.g., the corporation. We assume that they will behave as an individual would behave. We often speak of the firm, or the company, or the corporation as making certain types of decisions. Of course, this allusion is a fiction, since it is individuals in firms who make decisions and not firms themselves, although in some cases specific individuals may make them in the name of the firm. But for many purposes it is a convenient fiction. By contrast, organization theory goes behind this fiction, to examine intrafirm decision-making, and to see firm decisions as the outcome of the interaction of individual decisions. One purpose of the later sections of this book is to see if we can get additional insights into the behavior of such basic units as the firm by examining intraorganizational elements. But in the first six chapters (Part I) we shall set out briefly the essential elements of microeconomic theory from a rather elementary viewpoint.[1] This serves as an introduction to those parts of the book that deal with the application of some organizational ideas to economic analysis, and especially to a reconsideration of the theory of the firm.

Although conventional economics has usually not employed the phrase "decision theory," in fact microeconomic theory has always been about decision-making, and about the interaction between decisions, and between the activities that they imply. The following exposition will focus on the process of decision-making that is assumed within the standard theory.

A word of caution may be in order. We made the point in the preface that to understand a scheme of analysis it is not necessary to know all the variations possible, or all of the conceivable applications of the given analytical scheme. In the light of this view we shall not attempt to be exhaustive. Rather, we shall try to indicate the scheme of analysis by reviewing its essential elements.

THE ECONOMY AS A SYSTEM OF EXCHANGE DECISIONS

In microeconomic theory we visualize the economy as a vast and complex system of decision-makers and decisions. It is assumed implicitly that people act in accordance with their decisions. The decision-makers are buyers and sellers, or they may be classified as consumers, firms, or owners of agents of production. For the time being, in order to simplify our discussion, we assume that the government is not involved as an economic decision-making entity. Some of the decision-making entities are individuals; others, such as firms and households, are usually composed of groups of individuals, behaving, we assume, in the same way with respect to decision-making as if they were individuals.

Most individuals will normally act in more than one capacity. Thus, consumers may simultaneously own factors of production, such as their labor power or capital, and make decisions as to how to employ these productive agents. Firms must make decisions with respect to the disposition of the commodities they produce. To understand an economy we have to understand how such economic decisions are made and the consequences implied by the interaction of the decisions. It is these decisions and the activities they engender that determine economic events.

---

[1] The reader who is familiar with basic microeconomic theory, essentially the theory of price and output, may skip this part and begin with Chapter 7.

The scheme of decision-making assumed in microeconomic theory is in principle quite simple. The usual decision will either explicitly or implicitly involve an element of exchange. In other words, something usually has to be given up in order to get something. The thing given up can be looked upon as a cost, and the thing obtained leads to a benefit of some sort. The assumed object in making such decisions is to do as well as one possibly can, given the resources with which one starts. For example, the consumer is assumed to have a given income; his decision problem is to spend his income in such a way as to get as much satisfaction as possible from it. Sellers of factors of production will seek to sell them so that they get as much income as possible. Firms usually try to maximize the profits from their activities. In all cases the economic entities have to work with limited resources. The usual decision problem is to try to maximize something—for example, utility or profits—subject to the constraints of the limited resources available to the decision-maker, the technological information available to him, and the market conditions in which he operates. The sum total of such decisions determines what happens in an economy.

EXCHANGE DECISIONS

As we have already indicated, an economic decision is essentially an exchange decision. We have to balance what we can get against what we have to forgo in order to get it. Consider, for example, the problem of a consumer trying to purchase something in the market at a given price per unit. The more he buys the more he must spend for his purchase. His problem is to determine how many units to purchase, considering what he has to give up in order to make that purchase. There are two elements involved: how much he wants of a commodity at the given price—the demand element; and how much he is willing to give up in order to get it—the supply element. The same two elements are involved for the seller as for the buyer. To understand exchange, therefore, we have to analyze these elements of the situation from the viewpoint of the various parties involved in the exchange.

How should a rational individual make an exchange decision? If the price per unit is given, he has to answer the following question: Does what he gets as a result of the exchange give him greater satisfaction than having the thing that he would give up in order to make the exchange? If the answer is affirmative, then clearly it would be sensible for him to make the exchange. Further, he should be willing to continue to make exchanges as long as he can answer the question in the affirmative.

THE MARGINAL PRINCIPLE

A common approach is to assume that the consumer seeks to maximize utility, and that he endeavors to make his exchange decisions in such a way as to maximize his utility. By the utility of a commodity we have in mind the anticipated satisfaction that the commodity yields to the consumer. In normal cases it is assumed that more units of the commodity will yield more satisfaction in the aggregate, but that the amount of satisfaction (or utility) will increase in lesser proportion to an increase in the number of units consumed. We shall elaborate on these ideas later.

Closely connected with the concept of maximization is the marginal principle—a principle that is to some extent the very essence of the scheme of thought used in such matters. Consider the question, How many units of a commodity should a consumer buy? To answer this he should weigh the utility (or satisfaction) of a given quantity of the commodity against the utility to him of the money that he must give up for the commodity. For example, should he buy at least two units rather than one? Here he must weigh the *additional* utility that he puts on consuming two units rather than one against the additional utility lost from giving up the quantity of money necessary to obtain two units rather than one.

The *additional* utility obtained from consuming two units as against one unit is known as the marginal utility of the second unit. The same concept could be applied if we consider the utility obtained from consuming x units as against $x-1$ units. A similar idea is applicable with respect to the marginal utility of money. If the marginal utility of another unit, when one unit has been purchased, is greater than the marginal utility of the additional money outlay for the second unit, then it would be sensible for the consumer to purchase two units rather than one. Extending this notion, we may say that it would be sensible for him to purchase x units if the marginal utility of x units exceeds the marginal utility of the money given up for x units.

We have dealt here with utility or satisfaction, but we could express the same idea in terms of any type of marginal valuation placed on anything, whatever the index of valuation happens to be. It would make no difference whether our index were money, or income, or utility, or any other individual or social value. The appropriateness of the index, of course, depends on the problem.

At this juncture it is useful to introduce the law of diminishing marginal utility. This law states that as we consider consuming more units of a commodity the marginal utility associated with a greater number of units declines, *other things equal.* A commonplace experience is that the utility obtained from consuming, say, a seventh banana is likely to be less than from a sixth one, and certainly less than from the second one, or the first one consumed. We shall not try at this point to justify the law of diminishing marginal utility except to say that it is based on experience sufficiently pervasive as to be a useful tool of analysis. Whether or not it represents the universal experience is something we need not concern ourselves with at present.

Now let us return to our previous problem and assume that the law of diminishing marginal utility is applicable to the commodity in question. Let us say that the commodity is bananas. The consumer is faced with the problem of having to decide how many bananas to purchase. Suppose that money is subject to constant marginal utility. Once again, we will not attempt to defend this assumption because our main interest is to elucidate a process of thought and not to worry about realism. As we consider buying more and more bananas the marginal utility of bananas declines. Thus, the more bananas purchased, the lower their marginal utility. As we consider the prospect of purchasing more rather than fewer bananas, a point will be reached at which the marginal utility of an additional banana is no greater than the marginal utility of the money given up to obtain the last banana. When the number of bananas is such that the marginal utility of a banana is no greater than the marginal utility of the price of the banana, it will not pay to purchase any more. The point of equality[2] of the two marginal utilities yields the number of bananas to be purchased in order to maximize utility, for to buy more would mean giving up a greater utility in terms of the additional money outlay as against the utility of having an additional banana.

A slight variation on the above example will illustrate a closely related point. Suppose that we have a fixed budget to spend on both bananas and oranges, and that the price of a banana is equal to the price of an orange. Our decision problem is to pick a combination of bananas and oranges that we could obtain with our fixed budget that would give us the greatest satisfaction. Once again, we presume that the law of diminishing marginal utility is applicable to both commodities. The answer lies in allocating our fixed money budget among bananas and oranges in such a way that the marginal utility of bananas is equal to the marginal utility of oranges. That is to say, we should choose a combination of bananas and oranges that we could obtain by spending all of our budget so that the utility we derive from the last banana purchased is equal to that derived from the last orange. To see that this gives us the maximum utility let us suppose for a moment that we have picked a combination of bananas and oranges for which their marginal utilities are not equal; say, that the marginal utility of bananas is greater than the marginal utility of oranges. The question we raise is whether we can improve our position by giving up oranges and substituting bananas. We can clearly make such substitutions since we assumed at the outset that the price of an orange is equal to the price of a banana. Thus, if we purchase one less orange we can purchase one more banana. Such a procedure would clearly improve our situation because the added utility of an additional banana is greater than the utility of the orange forgone. Indeed, it would be sensible to keep shifting out of oranges and into bananas as long as the marginal utility for bananas falls and that of the oranges that we have left rises. Thus, the procedure of shifting will lead the marginal utilities of the bananas and oranges in our possession to approach each other. At some point, when they become equal, it will no longer pay to make any more shifts. We will then have reached a point where we have done as well as we possibly could.

---

[2] We assume that the commodity in question is completely divisible and that we deal with continuous variables.

The previous example is an illustration of the *equimarginal* principle. This is, perhaps, one of the central ways of thinking in microeconomic analysis, especially when we are concerned with the question of the allocation of resources. In general, then, if we want to allocate any fixed quantity of something among alternative uses so as to maximize their aggregate valuation, we should do it in such a way that our marginal valuations of what we get out of the resources are equal in each use. The previous examples are simply illustrations of this general approach. A great deal of microeconomic theory involves examining the implications of decision-making in which this principle is applied in one form or another.

DEMAND AND SUPPLY

A second basic approach in economics is to combine the potential decisions of a number of individuals in order to analyze an event or state of affairs. We will illustrate this method briefly by considering the analyses of exchange in the market through the use of demand and supply curves. By a market we have in mind all of the potential buyers and sellers in a given area that can possibly negotiate for the purchase or sale of a given commodity.

To analyze an event we have to separate it into the elements that contribute to its occurrence. This in turn depends on an understanding of what *can* occur—that is, on an appreciation of what the possibilities are. By separating the possible elements that contribute to an event or state of affairs, and by observing where these possibilities coincide, we obtain an explanation. How does this work out with respect to our understanding of an exchange event?

If there is only one possibility, the explanation is simple. The event is explained by our understanding that nothing else could occur. But usually a number of forces (or elements) contribute to the event, and it is the constellation of these that determines what happens. Different combinations of these forces bring about different results.

To see these ideas more clearly we use the concepts of relation and function. A relation is simply an association between different values of two variables. The price of a commodity and the quantity that a person is willing to purchase at that price are two variables, and some values of one will be associated with some values of the other. For example, at a very high price, say $5, only one unit may be purchased, but at a lower price of $1 ten units will be purchased. We have here two sets of associations: $5 and one unit, and $1 and ten units. There are usually many other associations of this type, all of which will tell us the connections between the amounts that people are willing to buy at different prices. The set of all such associations is a relation. When the relation is of such a kind that we can designate a dependent and independent variable, and if to each value of the independent variable (in this case the price) there is associated only *one* value of the dependent variable (in this case the quantity), we call such a relation a function.[3] In particular, the relation between the quantities of a commodity that people are willing to buy at various alternate prices is the demand relation. This is illustrated by the curve *D* in Figure 1.

Similar to the demand relation is the supply relation. This tells us the alternate quantities of a commodity that sellers would be willing to offer for sale at different prices. Figure 1 illustrates both

---

[3] The relations that we shall use in our illustration are functions.

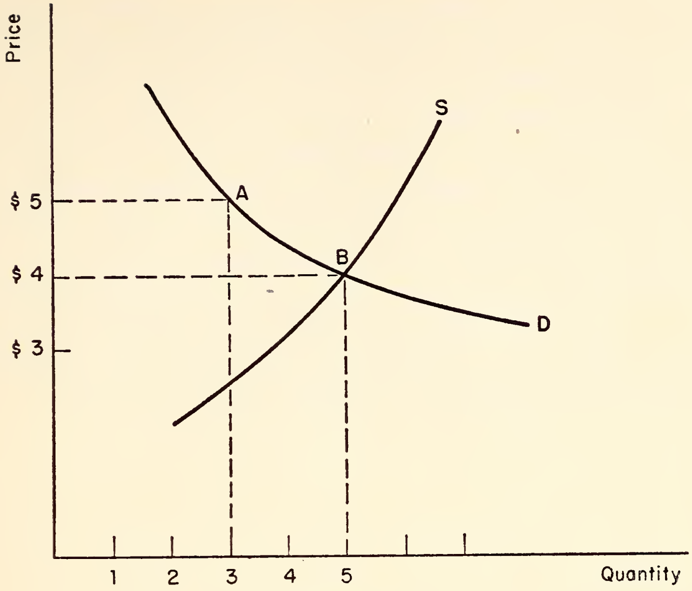


Figure 1.

the demand and the supply relations for a given commodity. The x-axis indicates the different quantities of the commodity while the y-axis represents the different prices at which they could possibly be bought and sold. The curve *D* in the diagram is an illustration of the demand relation, while the curve *S* illustrates the supply relation. Each point of *D* represents an association of a price and the quantity that people would be willing to buy at that price. Similarly, each point on S represents a price and an associated quantity that sellers are willing to offer for sale at that price.

Consider how we should interpret alternative points on the demand relation. Each point represents a possibility, a potential event that could occur if the other forces were of a certain kind. For instance, the point *A* associates $5 and three units demanded. If the seller were willing to supply three units at $5 and at no less than $5, then the buyer would purchase the three units at $5. Thus the point *A* represents a possible occurrence for buyers. Similarly, all other points on the demand curve represent possible occurrences for the buyers. By the same token, points on the supply curve in turn represent possible occurrences for the sellers. Of course, not all possibilities do in fact come about, because in most cases possibilities for buyers are not possibilities for sellers, and vice versa. The point *A* will not come about because at that price the suppliers are eager to sell many more than three units. However, at the point *B* we have a coincidence of the possibilities as seen from the demand side and from the supply side, that is, from both the buyers' point of view and the sellers' point of view. It is the coincidence of these possibilities that determines the event: the quantity that will be exchanged and the price at which it will be exchanged.

2

Demand Theory

UTILITY AND DEMAND

Our main aim in this chapter is to examine the demand element in exchange decisions.

We have already introduced the concept of utility and the law of diminishing marginal utility. We now consider how these ideas are connected with the demand relation, for a single individual and a single commodity. The first question to be answered is: How much of the commodity will the individual be willing to buy at different prices? To answer this question we can apply the principle expounded previously, namely, that the quantity he is willing to buy will be determined at the point at which the marginal utility of the good to be purchased is equal to the marginal utility of the money given up in order to make the purchase.

Suppose that at a price of one dollar the marginal utility of the tenth unit of the commodity is equal to the marginal utility of one dollar. We can try to answer the same question for a price of two dollars, a price of three dollars, and for all alternate prices and thereby obtain the demand relation. It is easy to see that in the normal case less will be purchased at a higher price than at a lower price. At a higher price the marginal utility of the expenditure necessary to purchase an additional unit of the commodity will be higher than at a lower price. Also the marginal utility of the good will be the higher, the smaller the number of units of the commodity purchased. Hence, the quantity at which the utility of the expenditure to obtain the marginal unit is equal to the utility of the marginal unit of the commodity will be the less, the higher the price, *other things equal.*

THE INDIFFERENCE CURVE APPROACH

An approach often used to explain consumer behavior is that of indifference curves. Among other things this approach enables us to examine the effect of income changes on consumer behavior, as well as the effect of a change in price of one commodity on the demand of another.

Suppose that a person has a given budget, and he has to spend all of it on two commodities. How will he react to price changes? How will he react to changes in the size of his budget?

Figure 2 shows a family of indifference curves. The x-axis shows alternate quantities of commodity X and the y-axis alternate quantities of commodity Y. Each point on an indifference curve represents a combination of X and Y. If X is oranges and Y is apples then a point on an indifference curve would represent a combination of apples and oranges. All points on the same indifference curve represent all the combinations (or bundles) of apples and oranges with respect to which the person is indifferent. That is, he feels as well off with one combination as with any other on the same indifference curve.

We have drawn the curves convex to the origin. A word must be said about the shape of the curves. Consider a movement along curve 1 in the direction of the arrow shown in the figure. A movement along the curve involves the substitution of X for Y in such a way as to leave the person equally well off. Now suppose that we invoke the law of diminishing marginal utility at this juncture. As

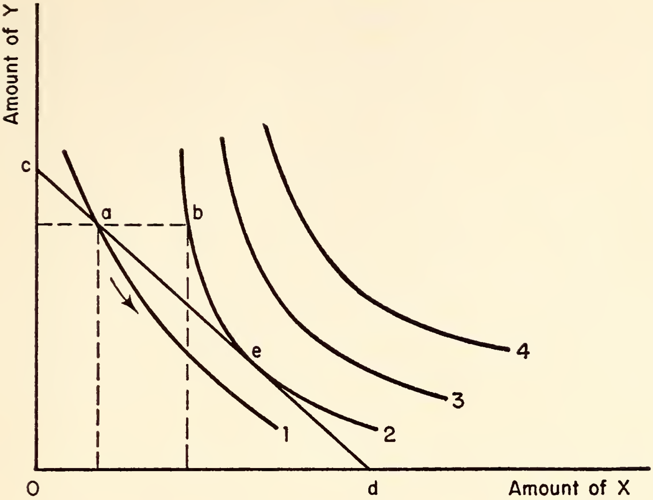


Figure 2.

we substitute X for Y we obtain more X and therefore the marginal utility for X declines. At the same time we give up Y and therefore have smaller quantities of Y; hence the marginal utility of Y rises. As a consequence, as we carry out our substitution in the direction of the arrow we have to substitute more and more X for given units of Y to obtain the same degree of satisfaction—that is, to remain at the same utility level. This explains the shape of an indifference curve.

If we assume, as is normally the case, that people prefer more of a commodity to less of it, then indifference curves farther from the origin will represent more preferred situations than those closer to the origin. Compare points *a* and *b* in Figure 2. Since *b* represents the same quantity of Y but a greater quantity of X than does point *a*, it follows that *b* is on a higher indifference curve than *a*, i.e., an indifference curve implying a higher utility or satisfaction level than *a*. The decision problem of the consumer is to find a bundle that is within his budget and at the same time is on the highest indifference curve that he can possibly get to, given his budget constraint.

We begin by supposing that the prices of the two commodities are given to the consumer in the market place; this is the usual situation faced by consumers in well-developed markets. They can purchase whatever amount of the commodity they desire at the stated price, but they cannot haggle over the price. Now given his budget (or income—we shall use these words interchangeably in this context) and the price of X, he can purchase a certain maximum amount of X shown by the point *d* in Figure 2. Similarly we assume that he can purchase a certain maximum amount of Y shown by the point *c* in the figure. Of course, he may choose not to spend all of his money on X, nor to spend all of it on Y. He may choose to purchase a combination of the two commodities. The line *cd* shows all combinations of the two commodities that he can purchase with his budget. The bundle that he will actually choose will be the bundle represented by the point *e* in the figure. Since it is the tangency point between the budget line and an indifference curve it is the highest indifference curve he can get to while still remaining within his budget.

That *e* is a bundle representing the best he can do with his budget can be seen from a slightly different point of view. Indifference curves reflect a man's tastes or preferences. Also, the extent to which a consumer is willing to substitute one commodity for another depends on the quantities of the two commodities possessed, (the law of diminishing marginal utility). Thus, as we go along an indifference curve substituting X for Y, we are willing to substitute more and more of X for the successive units of Y that we give up. Each point on an indifference curve represents a different rate of substitution of X for Y. In the market, at given prices, the market rate of substitution of X for Y is always the same. On the budget line, which reflects what the consumer can do in the market, the rate of substitution of X for Y is the same at all points. Now the point of tangency *e* is a point at which the consumer's preference rate of substitution of X for Y is the same as the market rate of substitution of X for Y.

But why is the point *e* better than any other point on the budget line? Consider, for example, the point *a* on the budget line *cd*. Suppose that at *e* he is willing to substitute one X for one Y, and that the market price for one X is the same as that for one Y. The point *a* represents less X and more Y than *e*. Therefore, at that point the consumer will value a unit of X more than a unit of Y. But since in the market X and Y are the same price it follows that by giving up a unit of Y he can obtain a unit of X, which he values more highly than Y. Clearly, it will pay him to make such a substitution of X for Y, and to continue to make such substitutions until the point *e* is reached, at which point the market rate of substitution is equal to his personal taste rate of substitution. By similar reasoning it can be seen that any other point on *cd* will be inferior to the point *e* for our consumer.

DERIVATION OF THE DEMAND RELATION

The indifference curve technique is also used to analyze the effects of a price change, and to derive the individual demand relation.

The consequences of a price change can be separated into two effects: an income effect and a substitution effect. If we have a given money income or money budget then a reduction in the price of a commodity, the prices of other goods remaining the same, implies an increase in *real* income. We can enjoy a higher level of satisfaction with our given money income and a lower price than we could previously at a higher price. Similarly, an increase in price implies a reduction in real income. But a change in real income usually leads us to change our consumption pattern. The extent to which we change consumption as a result of the change in real income (implied by a price change) is what we mean by the income effect. This is illustrated in Figure 3.

As in Figure 2 we begin with our budget line *cd*. A reduction in the price of X means that we can purchase more of X—say, *Og* of X. The lower price of X implies a new budget line *cg*, tangent to a new and higher indifference curve at the point *m*. Clearly, the total effect of the price change is the shift from the bundle of goods represented by the point *e* to that represented by the point *m*. The

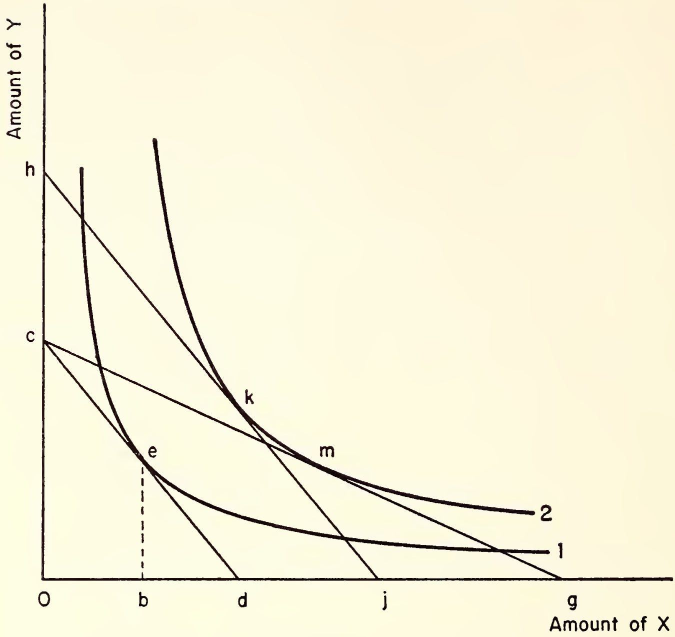


Figure 3.

point *m* is on a higher indifference curve, and the move to a higher indifference curve is equivalent to an increase in real income. Indeed, real income can be interpreted in terms of the level of satisfaction that can be reached with one's money income. Thus in order to show the real income change apart from the price change we have to consider what amount of money income (or money budget) would enable us to achieve the new satisfaction level. We observe that at the old prices, but at the new income level implied by the shift from *cd* to *hj*, we would purchase the commodity bundle *k*. We are now in a position to separate the two effects of the price change. The income effect is shown by the shift from *e* to *k*, while the substitution effect involves a further shift from *k* to *m*. In other words, if prices had not changed but our income had increased to enable us to get to the new satisfaction level, we would change our consumption pattern from *e* to *k*. But in fact relative prices have changed. Therefore at *k* the market rate of substitution is not the same as the consumer's preference rate of substitution. As a consequence a substitution of X for Y would take place, and we would shift from *k* to *m*. The actual shift is, of course, directly from *e* to *m*, but for purposes of analysis, in order to better understand the consequences of a price change, we separate the shift into its two elements, the income effect (from *e* to *k*) and the substitution effect (from *k* to *m*).

In Figure 3 the consequence of the price change is to increase the quantities consumed of both commodities. But it is important to note the possibility, which occurs in some rare cases, that the effect of a reduction in the price of a commodity may be to *decrease* the quantity purchased of that commodity. This can come about in the case of so-called inferior goods—goods that are consumed simply because one's income is low. Margarine is sometimes quoted as an example of an inferior good. When one's income rises one will readily substitute butter for margarine. If the inferior good forms a very significant proportion of the budget, and if the price of that good declines sufficiently, the consumer's real income may be increased to such an extent that he purchases less of the inferior good and employs the money left over to enjoy some of the superior substitute. The reader should find it easy to experiment with indifference maps of such shapes as to give the result we have just discussed. The important element is that the income effect may be in the same direction as the price change, and in some cases the income effect may be more important than the substitution effect. In such rare cases a decrease in price will be associated with a decrease in the quantity demanded.

Since we could use indifference curves and budget lines to show us the consequences of price change, we could also use this apparatus to derive the demand curve for the individual consumer for a single commodity. After all, a demand curve or demand relation is simply a set of points each associating the price of the commodity and the maximum quantity that would be purchased by the consumer at that price. By considering different prices for X, and the different budget line implied by each price considered (the price of Y remaining the same), we can obtain as many points on our demand curve as we want.

Let us return to Figure 3 and consider the derivation of the demand for commodity X. The budget line *cd* implies a given price for commodity X. The point *e* tells the quantity (in this case *Ob*) that the consumer would take at that price. Thus here are the two pieces of information necessary to determine a point on the demand curve. Similarly, the point *m* and the budget line *cg* give us the information for another point on the demand curve. By considering any number of alternate prices for X that we may wish, and by drawing in the appropriate budget line from the point *c* to the appropriate point on the abscissa, we can obtain the information for as many points on the demand curve as we may require. Such a procedure would give us the demand curve for X for a particular individual if his money income or money budget is given.

MARKET DEMAND

Many problems involve the demand relation for a group of individuals or for all consumers in a particular market. If the individual demand relations are independent of each other, i.e., if the quantity that one person is willing to buy at a given price is independent of how much any other person is willing to buy, then it is possible to aggregate the individual demand curves to obtain the demand curve for the group. What we seek is a series of points such that each point associates a price and the quantity that would be demanded by the group at that price. But the quantity that would be demanded by the group at any price is simply the sum of the quantities that would be demanded by the individuals composing the group. From each individual demand curve we know the quantity that would be taken by each individual at a given price, and by simply adding these quantities we obtain the quantity that would be taken by the group at that price. By doing this for all alternate prices that we may be interested in, we obtain the demand relation for the group.

ELASTICITY OF DEMAND

A concept often employed in connection with many relations in economics is that of elasticity. Elasticity is a measure of the relative change in one variable and the corresponding relative change in the other variable in the relation. E.g., if there is a change in price, what will be the proportionate change in quantity associated with it? Elasticity is a measure that will give us this information for small changes in price. We may represent elasticity in terms of the following ratios:

$$E = \frac{\text{Proportional change in quantity}}{\text{Proportional change in price}}$$

$$E = \frac{\frac{\Delta q}{q}}{\frac{\Delta p}{p}}$$

where q and p are the initial price and quantity, Δp is a small change in price, and Δq is the small change in quantity that results from the small change in price. Of course the elasticity usually differs for different points in the relation. It is a measure between one point in a relation and another point in its neighborhood.

Consider the elasticity of demand. If there is a small change in price, what will be the proportional change in the quantity demanded? Suppose that a 10 percent change in price leads to a 50 percent change in the quantity demanded; then the elasticity will clearly be greater than one. On the other hand, if a 10 percent change in price leads to a 2 percent change in the quantity demanded, the elasticity is less than one. An elasticity of less than one is usually referred to as an inelastic demand, while an elasticity greater than one is usually referred to as an elastic demand. In the normal case where the demand curve is negatively inclined, quantity changes will be in the opposite direction from price changes.[1]

It can be readily seen that elasticity of demand is related to the change in the monetary expenditure as a result of a price change. For example, if the elasticity is unity, the amount spent by a consumer will be the same after the price change as before. If the elasticity is greater than unity and there is a fall in price, the amount spent will be greater than before, and so on. Other relations of this kind can easily be worked out by the reader.

Too much importance should not be attached to the elasticity concept. It is simply a measure, and a piece of technical vocabulary that helps us to characterize portions of a demand curve or of a supply curve.

DEMAND FOR A FACTOR OF PRODUCTION

Thus far we have restricted our concept of demand to that for a consumer good. But we could just as easily carry out our analysis of demand for a factor of production used by a firm in the production of some commodity.

Consider an employer's demand for a particular type of labor. As before, the demand curve is thought of as a set of points that reflect a set of potential decisions. Each decision or potential decision must take into account what is to be given up as against what is to be gained. The variables here are the wage rate and the number of workers to be hired. Again the potential decisions are made by balancing the marginal valuation of a worker against the marginal cost—in this case the wage that must be paid for his labor. What is the marginal valuation of a worker? To the employing firm it is the value of the *additional* amount that hiring one more worker adds to production, usually referred to as the value of the marginal

---

[1] Because of the negative slope of the demand curve the elasticity of demand is often defined as follows:

$$E_D = -\frac{\Delta q}{q} \Big/ \frac{\Delta p}{p},$$

where E_D is the elasticity of demand, and the other symbols are defined as above.

product. If the value of the marginal product is greater than the wage, the firm will want to hire another worker. In this way the employer could determine the number of workers he would want to hire at each alternate wage and this would give the demand function for labor. A similar procedure could be used to determine the demand function for any other factor of production.

3

The Competitive Firm in the Short Run

THE PRODUCTION PROBLEM

The significant decisions that determine the supply of goods in a modern economy are the production decisions. We owe to Marshall[1] the analytical device of dividing production decisions into two types—short-run and long-run decisions. By the short-run decisions we have in mind the determination of how much to produce and in what way to produce the output of the firm on the assumption that the size of plant remains fixed. By the long-run period we have in view a decision period sufficiently long so that expansions and contractions of plant can be made. This chapter is devoted to the short-run supply problem.

Given the nature of the commodity, there are essentially two types of short-run production decisions: (1) how to produce a given quantity of the commodity and (2) what quantity of the commodity to produce.

---

[1] See Alfred Marshall, *Principles of Economics,* 8th ed., Macmillan and Company, London, 1924, Book V.

We consider each of these decision problems in turn.

THE FIRM IN A PERFECTLY COMPETITIVE MARKET

The basic production decisions in an economy are made by firms. We shall not try to define a firm in detail. This is a question to be examined later in this volume. For our purposes a firm is an entity that purchases agents of production in one set of markets and transforms them into commodities, which it then sells in another market. As already suggested, in microeconomic theory we assume that firms make decisions as though they were run by a single individual. Another basic assumption to be added here is that the firm will make its decisions in such a way as to maximize its expected profits.

In fact, firms make many other types of decisions in addition to production decisions. However, we are primarily interested here in considering production decisions. For this purpose it is convenient to begin with the simple case of the firm in perfectly competitive markets, i.e., the small firm that operates in perfectly competitive markets in both the areas in which it sells its product and those in which it buys the agents of production. Henceforth, we shall use the term "factor" to denote all of those things that a firm purchases for production purposes. Thus, labor time, raw materials, and supplies of various kinds are examples of factors of production.

Now, under perfect competition we assume that in each market there are a great many sellers and buyers, and that each buyer or seller is so small, compared to the volume of goods exchanged, that his activities can have exceedingly little influence either on the total supply or on the total demand. Hence any single buyer or seller can have no perceptible influence in determining the market price. Every buyer buys and every seller sells the amount he wishes to at the market price. In other words, the maximum quantities offered for sale or purchased by any single firm will be too small to have any practical effect on the total situation.

However, it is not quite true that the seller need sell at the market price. While he could not sell above the market price, since his potential customers will go elsewhere, he may if he wishes sell below the market price. It would be quite foolish for him to do so, however, since by the very nature of the market he could sell the quantity that he may wish to at the market price. Clearly, there is no reason for him to take a lower price than the market price. On the basis of such considerations we say that buyers and sellers in competitive markets are "price takers." They make their decisions on the basis that the market price is given to them, and that they could not change it.

THE PRODUCTION TECHNIQUE DECISION

The decisions of how much to produce and in what way to produce should be made simultaneously. However, since it is impossible to discuss such things simultaneously we separate the two problems. Given the nature of the commodity produced and assuming that we know how much we want to produce, our problem for the moment is to determine the way in which that quantity is to be produced.

If there is only one way the commodity could be produced, there is of course no decision problem. That technique must be used. However, usually there are a large variety of techniques available and hence the decision that has to be made is precisely what technique to employ. We are not now interested in the engineering question involved. For each technique that we can choose it is presumed that the engineering know-how is already there. Since the prices are given, and (since the quantity of the commodity to be produced is presumed known at this stage,) our problem boils down to choosing the technique that will enable us to produce that quantity at lowest cost.[2] Choosing a technique implies that we have to purchase various quantities of the factors of production in order to use that technique. Since we are not interested in the technological aspects of the problem we could define a technique for our purposes in terms of the combination of factors that we have to employ in order to produce the quantity in question.

---

[2] The need to produce the given output at lowest cost is implied by the desire to maximize profits.

To simplify matters at this stage, assume that there is a continuum of techniques.[3] This means that every combination of factors implies a technique, and that we can shift from one technique to another by substituting some factors of production for others. Since our aim is to produce the output at minimum cost, it follows that the prices of the factors will determine the technique that will enable us to produce the output at minimum cost.

One way to proceed would be to compute the cost of the bundle of factors necessary to produce a given output for each technique, and to simply pick the one associated with the lowest cost. But to see what is involved in having an optimal technique, consider how we should behave if we begin with an optimal technique and the price of one of the factors of production changes. The logic of the following analysis is very similar to that used in connection with the theory of demand in the previous chapters.

Suppose the price of factor A falls, the other factor prices remaining the same. We will be induced to shift from one technique to another by substituting the cheaper factor A for some other more expensive factor. But generally speaking, factors are imperfect substitutes for each other. We are likely to substitute first in those cases where it is easiest to do so. For example, if in our production process we substitute aluminum for lead, there may be some areas where it is quite easy to do so, and at the outset that is where we will make our substitution. As we increase the extent of substitution it becomes more difficult to make further substitutions. These can only be made in areas requiring a greater amount of aluminum to replace a given amount of lead. To return to the general case, as we substitute more of A for B, it usually becomes increasingly difficult to substitute A for B at the same rate. To continue such substitutions, we must substitute more and more of A for given units of B. In other words, the marginal rate of substitutability declines and at some points it will no longer pay for us to substitute A for B despite the lower price of A. To put the matter technically, a point will be reached at which the marginal technological rate of substitution is equal to the market rate of substitution. These terms have to be explained.

---

[3] An invention or discovery or new information would be characterized by a change in the continuum of techniques if it altered the alternative techniques available for adoption by the firm.

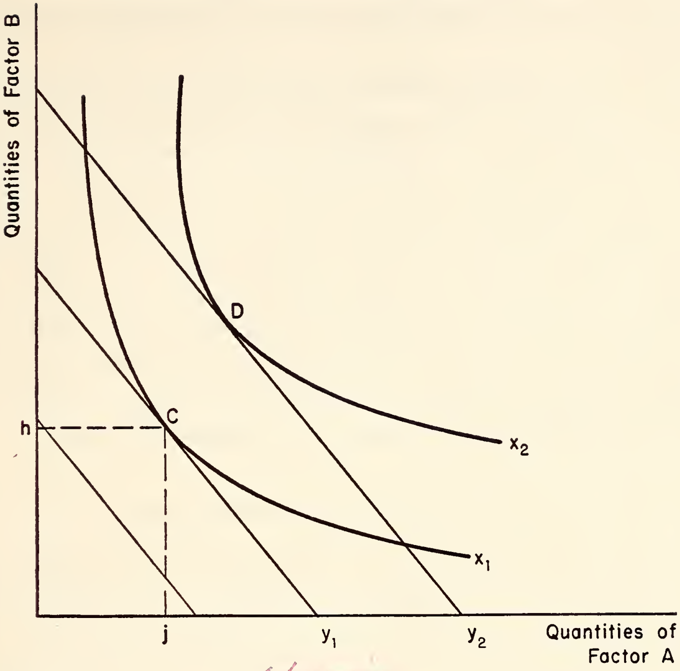
How much of one factor do we have to substitute for another *at the margin* to obtain the same output? It is important to note the qualifying phrase "at the margin." What we have in mind here is the ratio of substitution at the point that the substitution is being made, not the average ratio. Of course the marginal technical rate of substitution will continue to change as we substitute more and more of one factor for another.

By the market rate of substitution we mean merely the rate at which we can substitute one factor for another in the market place. Since the prices are assumed to be constant and given to us at the outset, the market rate of substitution will not change with the degree to which we carry on substitutions. It will remain the same for all degrees of substitution.

The foregoing ideas can be elucidated with the aid of Figure 4. The analysis parallels the indifference curve analysis in Chapter 2. The abscissa indicates alternate quantities of factor A, and the ordinate indicates alternate quantities of factor B. The curves in the figure represent production isoquants. Thus the curve x_1 denotes all the combinations of factors A and B which could produce the quantity of output x_1. Each point on a production isoquant represents a different technique, since each point represents a different combination of the two factors A and B. The slope at a point on a production isoquant denotes the marginal technical rate of substitution at that point.

In Figure 4 the production isoquants (x_1 and x_2) are drawn convex from the point of origin. This reflects the view that as we continue to substitute one factor for another substitution becomes more and more difficult. In other words, the marginal rate of substitutability declines.

(The lines y_1 *and* y_2 are outlay curves.) That is, they show the combination of the two factors A and B that could be purchased for a ordinate



abscissa

Figure 4.

given monetary outlay. For larger outlays larger quantities of both A and B could, of course, be obtained. These outlay curves are similar to the budget lines employed in the analysis of demand.

We recall now that the objective of the firm is to produce a given amount of output at minimum cost. Hence, for the quantity x_1 we want to get to as low an outlay curve as we possibly can. Figure 4 shows that to produce x_1 we can get to the outlay curve y_1, and this implies that we will use the technique indicated by the point C. This represents the use of h units of factor B and j units of factor A.

In a fashion similar to the above we can obtain an optimal technique, that is, one associated with minimum costs, for each quantity of output that we might want to produce. If we want to produce x_2, we will employ the technique indicated by the point D in the figure. Thus to each output we can associate an optimal technique, and a minimal cost for producing that output. We shall use this fact in the next section.

A and B are variable factors of production. Our scheme of analysis tells us how, for each output, we could determine how much of the variable factors A and B we want to employ and what the minimal cost of using A and B will be in order to produce the output under consideration. Indeed, if all of the production isoquants were included in Figure 4 we would have here all the information necessary to determine the output-total variable cost relation. That is, with the output x_1 we associate the (minimum) total variable cost y_1 necessary to produce it; with the output x_2 we associate the total variable cost y_2, and so on. This is the information necessary for the derivation of the four basic cost curves discussed below.

THE OPTIMAL OUTPUT DECISION

Our next task is to select the output that will maximize profits. To do this we have to consider two relations: the relation between output and revenue and the relation between output and costs. Once we know these two relations we can choose that output at which the difference between total revenue and total costs (i.e., total profits) is greatest. First we consider the cost-output relation and its variations.

Again it is useful to separate our analysis into short-run and long-run problems. This implies a separate consideration of short-run costs versus long-run costs. Short-run costs are the cost of those factors of production that could be changed with changes in output in the short period. There are presumably, in a short period, some costs that cannot be varied with changes in output. These are usually the costs of plant, some of the equipment, and some types of personnel, e.g., supervisory personnel. Therefore, in the short run we shall be concerned primarily with the output-variable cost relation and its derivatives.

The previous section has shown that we can associate with each quantity produced a certain minimum total variable cost for that output. But this implies that with each quantity we could also associate an average variable cost, since the average variable cost is simply the total variable cost divided by the output. We therefore have immediately a relationship which associates for each alternate quantity of output a certain average variable cost. This relationship is illustrated by the curve *AVC* in Figure 5.

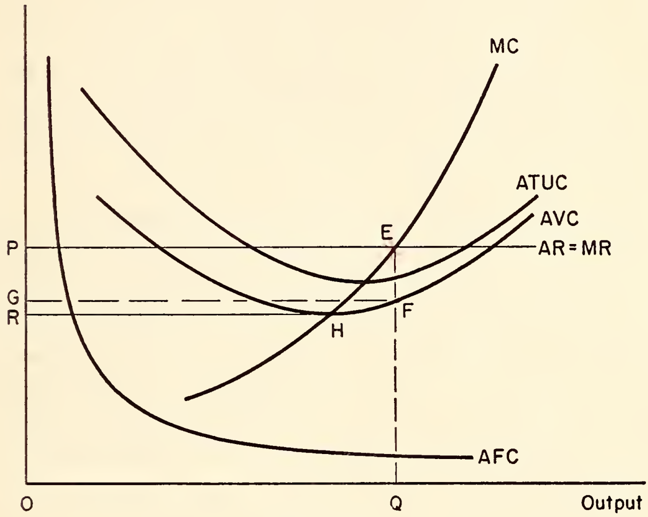


Figure 5.

Let us consider the shape of the average variable cost. As the curve is drawn in Figure 5, one portion of it is negatively inclined, reaches a low point at *H,* and then rises. Consider first the falling portion of the curve. Why should average variable costs fall with increases in the quantity of output? One possible reason is the potential advantages of specialization that could be achieved with increased output. Namely, when there are a number of variable factors of production as output is increased it is possible to choose a technique which is more specialized, and hence one in which the average costs of production are lower. We shall not go into detail here on this matter, since the nature and the advantages of specialization will be taken up in Chapters 7 and 8.

Beyond some point we should expect increases in costs as output rises. There exists at least one fixed factor in the short run. Thus as we add more and more of the variable factors, each unit of a variable factor has to be mixed or associated with, on the average, less and less of the fixed factors. We would expect at some point, therefore, that each unit of the variable factor has so little of the fixed factors to work with that the production process employed becomes less efficient and average costs rise. As a consequence we could argue that beyond some point we might expect rising average costs as we increased the quantity of output. Thus we end up with an average variable cost curve that is saucer shaped or U shaped.

The curve *MC* in Figure 5 represents the relationship between the additional (or *marginal*) cost of production and the quantity of output as one more unit of output is added. This is a relation derived from the output-total variable cost relationship. "Marginal costs" refers to the additional variable costs of producing one more unit of output. Now when average costs decline, the marginal (or additional) costs must be below or less than the average costs;[4] alternately, when average costs rise, the marginal costs per unit must be above average costs. Finally, when average costs neither fall nor rise, the marginal cost is the same as the average cost. The *MC* curve in Figure 5 is drawn in such a way as to reflect these general properties. It is shown to be below the *AVC* curve when the *AVC* curve is falling, above it when the *AVC* curve is rising, and at the same level on the *AVC* curve when the *AVC* curve is neither falling nor rising.

The two additional curves shown in the figure are very straightforward. *AFC* represents the relation between alternate quantities of output and the average fixed cost associated with those alternate quantities. The total fixed cost is, of course, the same regardless of the output. To obtain the average fixed cost, we merely divide the total fixed cost by the quantity involved. And it follows that with larger quantities the average fixed cost must decline. Indeed, the shape of the average fixed cost curve must be that of a rectangular hyperbola.

---

[4] Marginal costs below the average pull the average down, as it were, and vice versa.

The final curve in Figure 5 is the average total unit cost curve (*ATUC*). It represents the relation between alternate outputs and total unit costs. For each output the total unit cost is simply the sum of the average fixed cost and the average variable cost for that output.

THE REVENUE-OUTPUT RELATION

We have already argued that under competitive conditions the firm must be a price taker. Thus the revenue to be obtained from a single unit is simply the going price in the market place. As a consequence, the total revenue will rise in exact proportion with the quantity sold. The average revenue will, of course, be the same as the price and will not change with respect to the changes in the quantity sold. Hence, in Figure 5 the average revenue-output relation is denoted by the straight line *AR*.

We have already seen that when an average value does not change, its marginal value also does not change but remains the same as the average value. As a consequence, the marginal revenue-output relation (*MR*) could be denoted by the same curve as the *AR* curve. In deciding on the best quantity we will focus our attention on the marginal revenue curve and on the marginal cost curve. In the usual case the optimal output is determined at the point at which these two intersect.

In the normal case, where the marginal cost curve does not cut the marginal revenue curve from above, the most profitable (or least unprofitable) output is that associated with the point at which marginal revenue is equal to marginal cost,[5] i.e., the quantity *OQ* in Figure 5. To see that this is so, consider for a moment a quantity larger than *OQ*. At that quantity, marginal cost is greater than marginal revenue. Clearly, for the marginal units we are losing some of our profits per unit by the amount by which the marginal cost is larger than the marginal revenue. On the other hand, we might consider a quantity smaller than *OQ*. Here the marginal revenue is larger than the marginal cost. This implies that additional units of output would be profitable; it would pay to add to the output until the point is reached at which marginal revenue is no longer greater than marginal cost. Hence, the point at which profits are maximized is the point at which marginal revenue is equal to marginal cost.

---

[5] There are exceptions to this rule, especially in the long run; e.g., marginal revenue and marginal cost may not intersect at all, or there may be multiple intersections, in which case not all the intersections will imply optimal outputs. Cf. William Baumol, *Economic Dynamics*, 2nd ed., The Macmillan Company, New York, 1959, pp. 118–122.

QUASI-RENTS

To see the connection between the long run and the short run it is of special interest to focus attention on the difference between the price and the average *variable* cost. This difference is known as the quasi-rents per unit of output (*EF* in Figure 5). If we multiply the quasi-rents per unit by the quantity of output, we obtain the quasi-rents for the entire output. (This amount is shown by the figure *EFGP* in Figure 5.) Now these quasi-rents are the revenues less the variable cost. If the quasi-rents are larger than the fixed costs of operation, then, of course, there are some pure profits. But in any event, it is necessary during some periods that quasi-rents be larger than the fixed costs if the business is not to be an unprofitable one in the long run. However, it is not necessary that in every period this be the case.

In the long run the sum of the quasi-rents of all periods should at least equal the cost of the fixed assets involved, because fixed costs have to be covered sometime. More accurately, in making a fixed investment, which is a long-run decision, if the sum of the quasi-rents is greater than the costs of the investment, the investment *may* pay. One additional element must be taken into account and that is the interest rate. Other things equal, people usually feel that it is more desirable to have their income in the present period than in future periods. Hence, the quasi-rents of future periods have to be discounted for the interest earnings lost (i.e., forgone) by not having the revenue at the present time. We conclude therefore that if the value of the discounted stream of quasi-rents flowing from an investment is greater than, or at least equal to, the cost of the investment, then it is worth while to make that investment.

AN ASIDE ON EQUILIBRIUM AND ITS STABILITY

By the concept of equilibrium in economics, we have in mind a balance of forces or a state of rest. If the forces involved in determining a situation are in balance, and if there are no changes entering the situation from the outside, so to speak, then the system or the situation determined by the system will be one that is in a state of rest.[6] That is, it will be one which will repeat itself period after period.

A related way of viewing equilibrium (from the standpoint of microeconomics) is this: if an equilibrium exists, the decisions of different agents are all consistent with each other, and no one will attempt to alter his decisions as long as the environment remains the same. In disequilibrium some decisions will not be consistent with others, some people will try to change their decisions or will not be able to behave in accordance with their decisions, and hence the state of the system will not perpetuate itself in the absence of environmental changes.

An attribute of an equilibrium situation is its stability properties. An equilibrium position is stable if, when it is disturbed or displaced, a path of adjustment is set in motion that leads eventually to a return to the equilibrium position. To use a physical example, a small ball in a bowl will return to its equilibrium position in the center of the bowl if it is disturbed and momentarily placed elsewhere in the bowl. By contrast an unstable equilibrium is one that does not return to its original position when displaced.

The study of equilibrium positions and the elements that determine equilibrium conditions is usually referred to as static analysis.

---

[6] "Lack of change" would probably be more accurate than "rest," but it is a cumbersome phrase. "Rest" may imply a lack of activity, which is not what we have in mind. The magnitude of the activities are the same period after period in equilibrium.

The study of the stability of equilibrium requires dynamic analysis —that is, the study of the sequence of adjustment beginning with the disequilibrium position, a sequence in which events in one period influence events in the next. For some purposes it is convenient to distinguish between stability in the small and stability in the large. An equilibrium will be said to be stable in the small if it is stable for small displacements, even if it is not stable for large ones. Of course, an equilibrium that is stable in the large is stable for all displacements.

In the normal competitive market (in which the demand curve is negatively inclined and the supply curve positively inclined), the equilibrium will be stable if an excess demand (the quantity demanded less the quantity supplied at a given price) stimulates an increase in price, and if an excess supply stimulates a decrease in price. But more on this later. We return now to the specific question of short-run equilibrium.

In the short run we visualize an equilibrium situation as one that will persist if there are no short-run changes; but we would not expect it necessarily to persist in the long run. Since the difference between short run and long run involves changes in the fixed assets of the firm, then by short-run equilibrium we mean states of the system that last as long as there are no changes in the fixed assets of the firm. For the time being we shall concern ourselves with the short-run supply curve, which essentially may be looked at as a set of potential equilibrium points of a short-run nature.

THE SHORT-RUN SUPPLY RELATION

By the concept of the supply relation we have in mind, as already indicated, a set of points such that each point indicates the price and the quantity that would be offered at that price. Now we may also look at the supply relation as a set of potential equilibrium points. That is to say, each point on a supply curve *can* become an equilibrium point; it can become an output that will actually be supplied to buyers and will continue to be supplied if the demand situation is consistent with it and does not change. We will see how the demand situation determines whether a point is an equilibrium point or not a little later in this chapter.

First, we might consider the short-run supply curve of the firm. Let us return to Figure 5. Suppose that the price were *OP* as shown in Figure 5. Then the quantity that would be supplied by the firm would be *OQ*, indicated at the point *E* in the figure. Now let us consider another price, *OR*. The quantity supplied at this price is shown by the point *H* in Figure 5. Similarly, for every price we could associate with that price a quantity the firm would wish to supply. It would be, of course, the best quantity the firm would decide on, given the price situation. We might also observe that each of these quantity-price combinations falls on the marginal cost curve of the firm, or that portion of it above the point *H*.

Here then is a coincidence between the marginal cost curve of the firm and its supply curve. However, we have to make one qualification. It would be unwise for a firm to produce any output if it could not cover at least its out-of-pocket costs. Thus if the price in the market were below the lowest possible average variable cost, the firm would not produce anything at all. Looking at Figure 5 we see that if the price were below *OR* the firm would be wise to produce no output whatsoever. Hence, it is only that portion of the marginal cost curve above the point *H* that becomes the short-run supply curve of the firm. By knowing a competitive firm's marginal cost and by knowing the low point of the firm's average variable cost curve, we can determine that firm's short-run supply relation.[7]

Now let us consider the supply relation for all of the firms in the industry. Clearly, this is simply the aggregate of the appropriate quantities supplied at alternate prices for each firm. Suppose there are five firms of the kind shown in Figure 5. Then at the price *OP* the quantities supplied would be *OQ* for the first firm, *OQ* for the second firm, and so on for all five firms. By adding the quantities supplied for all five firms at that price we could obtain one point on the industry supply curve. Similarly we could obtain all the other points on the industry supply curve. Even if the short-run marginal cost curves of each of the firms were different, we could still obtain, using the same procedure, the supply curve for each firm, and, by adding the quantities connected with the appropriate prices for all of the firms, the supply relation for the industry as a whole. This yields an industry supply curve which we will want to consider in connection with the demand curve of the industry.

---

[7] Assuming, of course, that the firm maximized profits in the sense that we have used that term.

We can now confront the industry supply curve with the industry demand curve and obtain the equilibrium price and equilibrium quantity. This is illustrated in Figure 6. *SS* is the short-run supply curve. D_1 is the relevant demand relation. We observe that OP_1 is the equilibrium price and OQ_1 is the equilibrium quantity. OP_1 is the equilibrium price in that it is the price at which the quantity supplied, OQ_1, is exactly equal to the quantity demanded, OQ_1. In

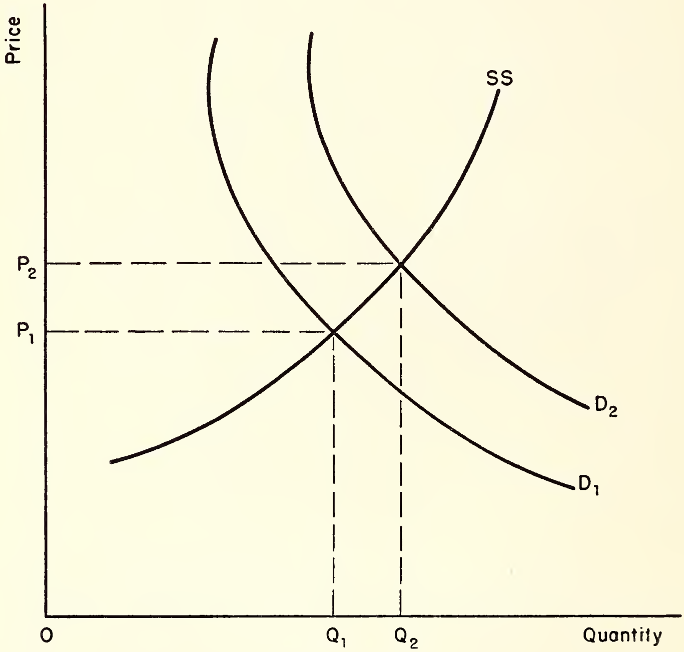


Figure 6.

other words, it is a price which will persist as long as the demand curve remains the same. At that price there is no incentive to either increase the price or decrease the price under existing conditions.

We can observe this lack of incentive for change if we consider, for the moment, a price other than OP_1. Suppose the price were higher than OP_1. In that case the quantity supplied would be larger than the quantity demanded. This would normally lead to excess suppliers unable to find enough customers, and in their attempts to sell more of the commodity than is wanted, they would try to lower the price. Clearly this is a state of affairs in which change is engendered. Similarly at a price lower than OP_1 there would be an excess demand. Customers who could not obtain any of the commodity at the price below OP_1 would then be willing to offer more for the commodity and in doing so raise the price. Once again this is a state that engenders change. As a consequence we see that OP_1 is the equilibrium price, OQ_1 the equilibrium quantity, and both combined gives us an equilibrium point, a point which would persist in the absence of exogenous changes.

Now we also observe that if the demand curve were different from D_1—say, that shown by the curve D_2, we would get a different equilibrium price. In the figure this is shown by OP_2, and in association with it is a new equilibrium quantity, OQ_2. Similarly with any given change in the demand curve we would get a new equilibrium price and quantity point. This illustrates our previous assertion that we may look upon the short-run supply curve as a collection of potential equilibrium points.

SUMMARY

We have seen that the prices of the factors of production determine the technique that is to be chosen. By technique we mean the combination of the factors to be used in producing a given quantity of the output. We have also observed that a change in factor prices will lead to the substitution of the relatively less expensive factors for the relatively more expensive ones. Also, the logic of choosing an optimal technique is similar to the logic of a consumer's choosing the best combination of consumer goods out of a given income.

The demand for a factor of production depends on that factor's productivity. Specifically we have seen (in Chapter 2) that it depends on the value of the marginal product of that factor. We would add units of a factor as long as the factor price is less than the value of the marginal product, and the optimal quantity of a factor to be employed is determined at the point at which the market price of the factor is equal to the value of its marginal product.

In addition to the demand relation, a crucial element in the determination of the optimal output is the cost-output relationship. Cost depends on the quantity produced. Up to a point we normally expect average costs to decline because of the advantages of specialization that could be obtained with greater outputs. Beyond some point we would find that with each additional output the variable factors are crowding on the fixed factors, that is, they have less of the fixed factors to work with on the average. In view of both of these considerations we expect the average total unit cost curve and the average variable unit cost curve to be U shaped.

The optimal quantity is that quantity which yields maximum profits. In the normal case it will be the quantity of output that is associated with the point at which the market price is equal to the marginal cost of production. For each alternate market price we obtain a different optimal quantity of output. Hence, we have here a collection of points of the following kind: each point in the set associates a price and the optimal quantity that would be produced and offered for sale at that price. This set of points is the short-run supply curve of the firm. We have also seen that we could aggregate all of the individual short-run supply curves for all of the firms in the industry and obtain the short-run supply curve for the industry. Finally, if we confront the short-run supply curve for the industry with the demand curve for the commodity, and if these curves intersect, then the point of intersection, assuming it intersects only once, determines the short-run equilibrium price and quantity for that market.

4

The Competitive Economy in the Long Run

THE decisions to enter or leave an industry and to expand or contract a firm, and the determination of the size of the firm, are usually considered long-run decisions. As before we assume at this stage of our discussion that the firm seeks to maximize profits, and by the optimal size of firm we have in mind the one that maximizes profits.

Now, the short-run equilibrium state may stimulate expansion or contraction. Whether it does so or not will depend on whether the short-run equilibrium is one in which there are profits or in which there are losses. In part, the long-run problem is to determine, when an inducement for expansion or contraction exists, the extent of the expansion or contraction that is most profitable. This will depend on two elements: the anticipated average price of the commodity in the long run, and the long-run average cost curves. Therefore, we shall focus our attention on the nature and possible shapes of the long-run average cost curves.

Once again we shall see that the logic of decision-making here is the same as before. The decision-maker is assumed to have information or definite expectations about the average market price in the long run, information about the nature of the long-run costs. (Here too the firm is a price taker.) The problem is to pick an ideal output for the long run, which in turn implies an ideal size of plant to produce that output.

LONG-RUN COSTS

By costs the economist usually has in mind all of the payments to all of the factors of production necessary to produce the output. They include not only such conventional factors as labor and raw materials but also interest on capital utilized in the production process, as well as payment for the managerial and entrepreneurial efforts and skills that may be employed by the owner of the business.

Suppose for a moment that the owner of the business also manages it. What is the appropriate cost with respect to the managerial effort? The appropriate criterion here is to use *opportunity* cost. That is, we try to answer the question, What could this owner earn if he were employed as a manager at the best opportunity elsewhere? It is his potential earnings that become the imputed cost for the owner's managerial services. The same will be true for all other costs. As a consequence it can be seen that if there is any revenue above all costs (that is, if there are any profits in the economist's sense of the term), then there should be an inducement for expansion, as long as costs do not rise precipitously with increases in the scale of operations.

LONG-RUN COST CURVES

As in the short run, the long-run cost curve is a cost-output relation. In other words, with each alternate quantity considered we want to associate the lowest cost per unit of producing that quantity, on the assumption that the decision period we have in mind is sufficiently long so that all factors are variable. Each quantity of output implies a certain firm size and a combination of the other factors of production that will enable us to produce that quantity at the lowest cost. Three broad basic possibilities may be considered: the long-run constant cost case, the long-run decreasing cost case, and the long-run increasing cost case. All of these are illustrated in Figure 7.

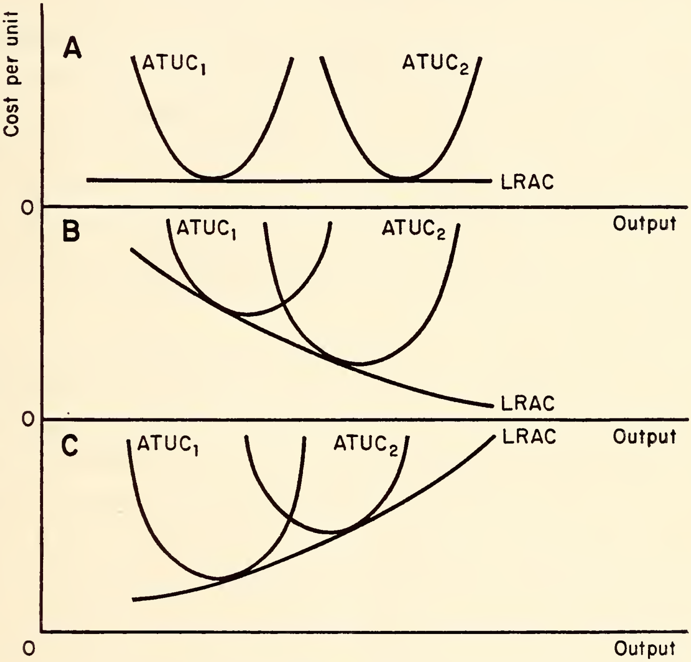


Figure 7.

The long-run constant case is perhaps the simplest. Let us look at it first, in section A of Figure 7. The curves marked $ATUC_1$ and $ATUC_2$, etc., are alternate average total unit cost curves of the type we considered in the preceding chapter—they are the short-run average cost curves of the firm. Each curve implies a different firm size. We assume as usual that there are infinite gradations possible and consequently an infinite number of such short-run *ATUC* curves. We observe that for each short-run cost there is a lowest cost point, and the curve marked *LRAC* (the long-run average cost curve) is the locus of the low points on the short-run cost curves. This states that the minimum cost of production per unit is the same for every firm size, i.e., it is the same for the small firm as it would be for a large one.

Figure 7B illustrates decreasing long-run average costs. As before, the curves marked $ATUC_1$ and $ATUC_2$, etc., are the short-run average total unit cost curves for larger and larger scales of operations. If we had an infinite number of these, we could draw a line which would be tangent to all of them, and we would obtain the long-run average cost curve, represented in the figure by *LRAC*. Now there are a few things we might observe about the negatively inclined long-run average cost curve. First, as before, each point on the long-run cost curve is also a point on a particular short-run average cost curve. It is the point that produces the particular quantity at lowest cost per unit.

It is of special interest to note that the points on the long-run average cost curves are *not* the lowest points on their respective short-run curves. The reason is that with economies of scale it will usually pay to build a slightly larger plant and underutilize it somewhat, rather than to build a smaller plant and utilize it at that plant's optimal capacity. The important thing to remember is that we are not interested in producing our output at any specific plant's optimal capacity. What we are interested in is producing the quantity involved at the lowest cost of production for that quantity, even if, as will be the case here, this requires building a larger plant and underutilizing it, as will quite commonly be the case where there are indivisibilities of some sort. It will often be possible, that is, to shift from a smaller machine to a larger machine, underutilize the larger machine, and still achieve a saving in cost.

Finally, in Figure 7C, we consider the possibility of increasing long-run average costs. As before, the *ATUC* curves are the short-run cost curves, and the curve marked *LRAC* is the long-run average cost curve obtained by drawing a line tangent to each of the short-run average cost curves. Again, each point on the long-run curve is simultaneously a point on a short-run curve.

Here we observe the opposite situation to the one examined with respect to decreasing costs: it will pay in the increasing cost case to build a plant smaller than the one that could produce the output at the plant's minimum cost and to overutilize it. Thus, the points of tangency between the long-run and short-run cost curves are drawn in such a way that they are to the right of the low points on the short-run cost curves.

Often textbook presentations of long-run average cost curves show them to be U shaped (see Figure 8). Judging by the prominence that such curves receive in textbook treatments, they represent the general case. But it is not too clear what economists, by and large, believe on this matter. One may perhaps take an eclectic viewpoint and argue that we need not commit ourselves to any given shape but may simply present the main alternatives. What is indisputable perhaps is the fact that different shapes of the long-run average cost curve do lead to very different conclusions with respect to analysis and policy. Further, it is clear that we must understand the underlying influences that determine increasing and decreasing returns, and their counterparts decreasing and increasing costs, if we are to appreciate the likely shape of the average cost curve under various circumstances.

LONG-RUN COMPETITIVE EQUILIBRIUM

The U-shaped *LRAC* curve is the one usually associated with competitive conditions in the long run. The shape of this curve is based on the belief that generally, or at least often, we will have decreasing costs up to some scale of operations, but that beyond some point the advantages of specialization[1] become rather small, and the disadvantages involved in the complexity of organizing resources on a large scale become more significant. As a consequence decreases in unit costs give way to increases in unit costs with increases in the scale of operation.

The central problem for the firm, in the long run, is to pick its ideal size, i.e., to determine its scale of operations. The logic of the decision is similar to that involved in the determination of the ideal output in the short run. To illustrate this let us look at Figure 8. The curve marked *LRAC* is the U-shaped long-run average cost curve. To every average cost curve we can associate a marginal cost curve; the curve marked *LRMC* in the figure is the appropriate long-run marginal cost curve. Once we have a marginal cost curve and an anticipated price we can determine an ideal output. Looking at Figure 8 we see that if the price is OP_1 then the ideal long-run output, determined at the intersection of the price line and the long-run marginal cost curve, is OQ_1. The point *F* on *LRAC*, related to the quantity OQ_1, is also a point on a short-run cost, which in turn implies the scale of operations and the firm size.

---

[1] On the advantages of specialization see Chapter 8.

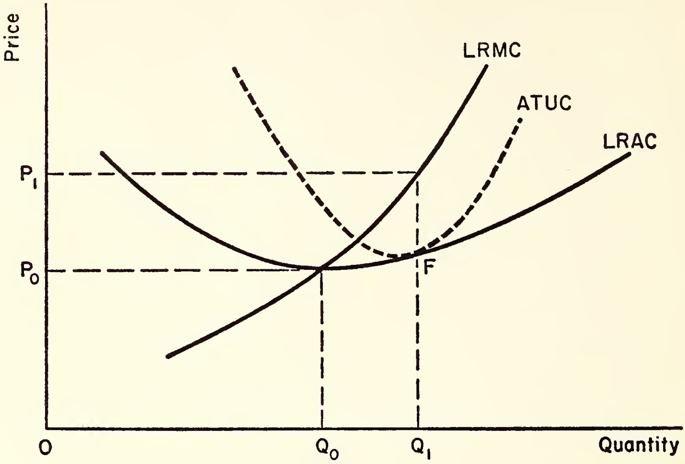


Figure 8.

By perfect competition in the long run we have in mind a state of affairs in which new firms can enter the industry under the same conditions as the firms already in the industry. That is, potential firms have the same access, and on the same terms, to technical information, market information, and the factors of production as existing firms. Hence new firms have the same long-run cost relations as old ones.

Now, returning to Figure 8, if the price were actually OP_1 we see that there would be absolute profits. The existence of positive profits would induce entry of new firms, and as a consequence increase in supply, which in turn would normally lead to a fall in price. Clearly OP_1 is not a long-run equilibrium price, nor is OQ_1 a long-run equilibrium output. Similarly, a price below OP_0 would also *not* be an equilibrium point. Such a price would involve losses, which in turn would induce exits from the industry, a reduction of supply, and eventually a rise in price. The only equilibrium situation is the one at which the price is OP_0. In this case the quantity of output and the related scale of operations is as indicated by OQ_0.

We might observe some of the properties of the equilibrium position. At this point the price is equal to the cost of production. Profits are zero. The output is produced at the lowest point on the long-run average cost curve, as well as on the lowest point on the related short-run average cost curve. Finally, these properties hold for all firms in the industry.

But it is important to remember that this last conclusion is due to the particular assumptions associated with perfect competition, some of which we will relax later. That is, we assumed that all firms have equal access to all relevant markets, and that the markets, for all factors, are purely competitive. This means that any factor of production could be obtained by any particular firm under the same conditions as it could be obtained by any other firm. In any market no firm would have an advantage over any other one. Therefore, for each quantity of output the optimal combination of factors would be the same for all firms. Also, the minimum cost for producing any quantity of output would be the same for all firms, which in turn, of course, implies that the long-run average cost curves are also the same for all firms. Thus, the perfectly competitive equilibrium scale of operations and the competitive equilibrium size would similarly be the same for all firms.

From the same point of view examine the possibility of long-run constant cost curves. (See Figure 9.) As before, the line *LRAC* is the long-run average cost curve for the typical firm. Now if the price is the same as the constant long-run average cost curve, then

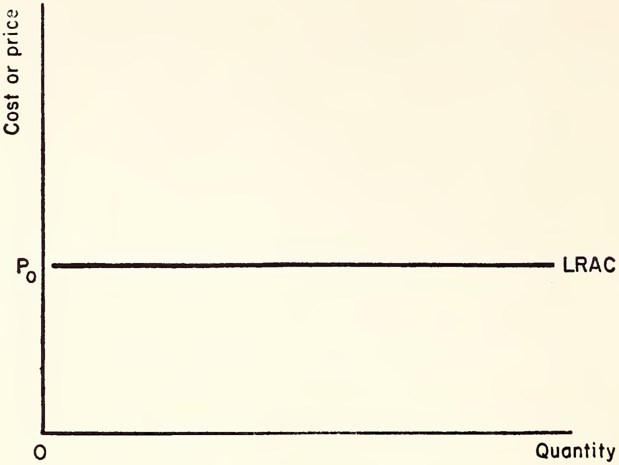


Figure 9.

we would not expect any inducement for firms to either expand or contract. However, we must observe that in this case there is no necessity for all firms to be the same size. It is quite possible under this situation for some firms to be small and for others to be large. Indeed there is no way of knowing in this situation how large any specific firm will be. Should the price be above average cost per unit, of course, there would be an inducement to expand, but nothing in this situation tells us whether some firms will not expand more rapidly than others. As a consequence there is an indeterminacy in this case with respect to firm size. However, if we can assume at the outset that each firm is sufficiently small and that the price is OP_0, then this is a situation compatible with long-run equilibrium.

AN ASIDE ON DECREASING COSTS

We shall not consider the decreasing cost case in detail because decreasing costs in the long run are incompatible with competition. It will be recalled that we defined competition as a state of affairs in which each firm is too small to influence market price by its behavior. However, if firms do have long-run decreasing costs, it will pay them to expand beyond some small size—to a size, that is, where price and output decisions can significantly influence the market price. Since this case is incompatible with the persistence of competitive equilibrium, it is best to consider it in connection with monopoly, oligopoly, and other topics that we shall take up later in which decreasing costs play a role.

UNIQUE FACTORS, COSTS, AND FIRM SIZE

Our analysis of competitive equilibrium was based on the assumption that each firm had equal access to all factors of production, and that all units of factors of a given type were the same for all firms. Let us now drop this assumption and see what happens. Suppose that firm A has a unique factor, say an unusually efficient manager. What is the result? First, the fact that the manager is more efficient means that firm A will have lower costs than any of the other firms. But the difference in costs between firm A and other firms is due entirely to the efficiency of the manager. As a consequence, the manager's salary should include this difference in cost between his firm and the other firms. In other words, the opportunity cost of the manager is equal to the manager's ordinary salary plus the difference of costs between firm A and the other firms. Another way of looking at it is to see that the manager of firm A could move to any other firm and thereby reduce the costs of the firm he moves to and raise the costs of firm A. Therefore, if the manager receives a salary equal to his opportunity costs, it will absorb the difference in costs between firm A and other firms. Thus, the actual costs of all firms will be the same, inclusive of the differential payment to the differential efficiencies of different factors employed. Usually we call the payment to a fixed factor (that is, one whose supply cannot be increased by an increase in price) a rent. Hence, we would say that the costs will be the same for all firms inclusive of rent.

But it is important to observe that even if the costs per unit of output are the same for all firms inclusive of rent when there are factors of different efficiencies employed, the size of firm need not be the same. For example, our more efficient manager of firm A may be able to organize a much larger number of individuals and run a much larger firm than the managers of other firms. The optimal scale of operations for the firm with the more efficient manager may be significantly different from the optimal scale of operations for the other managers. Firm A, then, may be larger than firm B, although their minimal costs per unit, inclusive of rent, will be the same in the long-run competitive equilibrium state.

INTERNAL AND EXTERNAL ECONOMIES OF SCALE

In the discussion of costs thus far, we have assumed that all economies of scale were internal economies. However, economies may accrue to every firm in an industry because the industry itself becomes larger. For example, a raw material may become cheaper as an industry becomes larger and as the demand for raw material increases and the cost of producing it falls. Now, the relative competitive position of every firm in that industry remains the same, since the cost for every firm in the industry falls in precisely the same way. However, the cost of the production per unit, as seen from the point of view of the industry as a whole, falls. A second example we might consider is that delivery costs per unit may decline as the demand for the commodity within a given geographical area expands, and as more units of products are delivered per unit of distance.

The importance of the distinction between internal and external economies lies, in part, in the fact that internal economies of scale are not compatible with competition, but external economies of scale are. For example, the supply price in an industry may fall as demand expands, but costs per unit may increase to any firm if that firm attempts to expand. In brief, external economies do not stimulate expansion in the size of any individual firm, although they do decrease costs for all firms with increases in output in the industry as a whole.

THE LONG-RUN SUPPLY CURVE UNDER PERFECT COMPETITION

In the short-run situation we were able to derive the short-run supply curve simply by adding the supply curves of all of the individual firms. This cannot be done for the long-run situation.

Consider the usual competitive case of U-shaped long-run average cost curves for each firm. We have seen that with an increase in demand the industry could expand in two ways, either by expansion of individual firms or by the entrance of new firms into the industry. In the competitive equilibrium state in the long run the firms in the industry will already be producing at their most efficient outputs. Therefore, the long-run expansion in output will take place through the entrance of new firms into the industry. Eventually the new firms will, in their turn, produce at their most efficient outputs. Although for each firm, apart from the others, the supply curve is its marginal cost curve (and this means a rising supply price), for the industry as a whole the supply curve is infinitely elastic;[2] that is, it is a line horizontal to the quantity axis, because as the firms enter the industry they produce as efficiently as the existing firms. Therefore, cost per unit does not rise with increases in output in the industry as a whole.

Now let us consider the possibility of external diseconomies of scale. In this case the result will be a long-run rising supply curve for the industry as a whole. As the industry expands, the demand for the factors of production rise, and factor prices will rise accordingly. But the increase in cost will be the same for all firms, and each firm will therefore continue to produce at its low point, although, since all of the cost curves of all of the firms must rise, these low points will now reflect higher cost than before. Therefore, from the industry point of view we will have a rising supply price. But note once again that there is no unique relationship between the marginal cost curves of the individual firms and the supply curve of the industry in the long run.

---

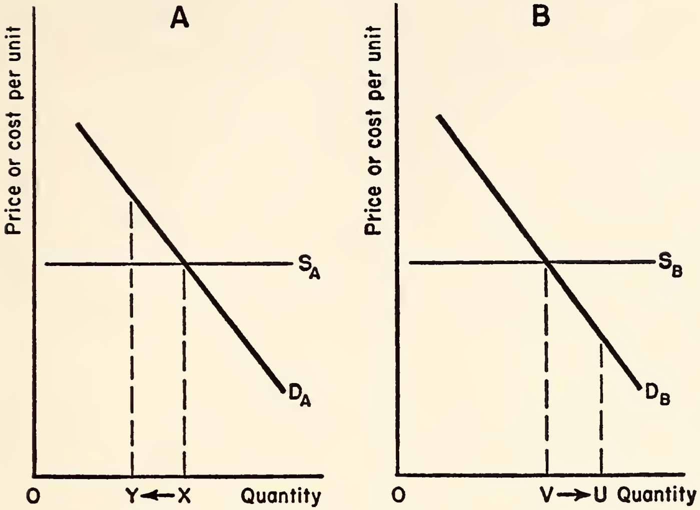
[2] Abstracting from the possibility of external economies or diseconomies of scale.

SOME CHARACTERISTICS OF GENERAL LONG-RUN COMPETITIVE EQUILIBRIUM

Thus far we have considered only one industry at a time. How about an entire economy made up of perfectly competitive industries? For the time being we will abstract from external economies of scale. We assume that every industry in our model competitive economy is in long-run equilibrium. Now, the state of general long-run equilibrium has certain desirable features.

First, under general competitive equilibrium the output produced is the best possible from the point of view of the consumers. In each industry the demand for the commodity is equal to its cost. Also, as we found in our consideration of the theory of demand, when a consumer is in equilibrium his marginal valuation of each commodity purchased is equal to the price that he paid for it.

But it is important to note that any output other than an equilibrium output is one for which, for some consumers at least, the marginal valuation placed on some commodities is greater than the cost in some industries, and less than the cost in others. This situation is non-optimal from the consumers' viewpoint. To see this point consider Figure 10. In the initial situation the two industries, A and B, are in equilibrium. We observe that the equilibrium output of industry A is a quantity *X*, and the equilibrium output of industry B is a quantity *V*. Now suppose that there is a shift of resources out of industry A and into industry B. This decreases the output in A and increases the output in B. The figure shows an output shift from *X* to *Y* in A and from *V* to *U* in B. In industry A the demand price is above the cost per unit of the commodity and in industry B the demand price is below the cost per unit. This implies that in A the marginal valuation of some consumers is greater than the cost, and that in B the marginal valuation is less than the cost. Clearly consumers will not want *U* units of output of commodity B at a price equal to average cost. The reason is that at *U* units they would be paying more for the marginal units of the commodity than the price at which they value it. If consumers do



S_A and S_B are the long run supply curves.
D_A and D_B are the demand curves.

Figure 10.

not buy the *VU* units of commodity B, then the resources used to produce them are clearly wasted. If on the other hand producers choose to produce only *V* units of B, because that is all that they could sell at a price to cover their costs, then some resources will be unemployed, which is also wasteful.

Obviously a shift of resources back from industry B to industry A will improve the situation. Mainly it will lead to an output that will be superior for consumers. For example, those who consume commodity A now find that they could purchase more of it, and they will wish to do so, since their marginal valuation for the commodity is greater than the cost. Hence, in the equilibrium state we see that consumers of A and B are better off than in the disequilibrium state. A similar type of analysis would show that the same would be true for any disequilibrium state. Thus, a general state of long-run equilibrium is one that yields an optimal output of the different commodities from the consumers' point of view.

A second desirable charactertistic of long-run equilibrium is that each commodity is produced at its minimal cost—not only at the minimal cost for that output, but also at the lowest point on the long-run average cost curve, and hence at the lowest cost at which it could possibly be produced. Furthermore, since price is equal to cost of production, it is clear that consumers benefit from the low cost of production. Also, as we have seen in Chapter 3, producing the output at minimal cost implies that the most economic technique of production is employed. In other words, each producer is induced to employ precisely that combination of factors that is most economical.

We know that in competitive equilibrium each factor of production receives a payment exactly equal to its contribution to the output of the economy. The payment to any factor is equal to the value of its marginal product. But the marginal product of a factor of production is precisely the additional amount that that factor contributes to output, and hence the value of the marginal product is equal to the return of that factor. Thus, for example, the wages of each type of labor will be equal to the value of the marginal product of that type of labor.

Finally, we observe that in the state of competitive equilibrium the factors are so allocated, both between firms and between industries, as to yield the greatest output. In other words, in this case we get an optimal allocation of each of the factors of production.

Here we repeat to some extent the argument used in the previous paragraph. In each industry the cost of the factor is equal to its price, which in turn is equal to the value of the marginal product of that factor. As usual, we assume diminishing marginal productivity with increases in use of each factor. Now it can readily be seen that any allocation other than an equilibrium allocation will lead to an inferior result. To see this we begin with an equilibrium state; then suppose we shift some units of a factor from one firm to another, say from firm H to firm J, other things remaining the same. (See Figure 11.) The result of the shift will be as follows: In the firm out of which the units of the factor were shifted (firm H) the price of the factor will be less than the value of the factor's marginal

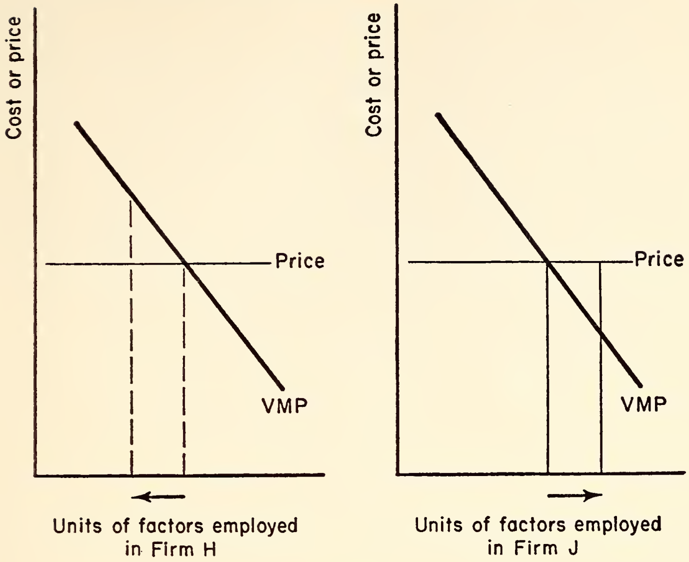


Figure 11.

product. In the firm to which the units of the factor were shifted (firm J) the price of the factor will be greater than the value of its marginal product. We note especially that after the shift the value of the marginal product is relatively high in firm H and relatively low in firm J. Now it should be clear that, if we shift some of the factor back from the firm in which the value of the factor's marginal product is low to the one in which the value of the factor's marginal product is high, we must increase the value of total output.

After all, what we mean when we say that the value of marginal product is low is that in that particular firm (and use) the marginal units of the factor of production add little by their work to the value of total output; and when it is high we mean that they add a great deal to the value of total output. And the value of total output is to be viewed in terms of what the firm can sell the output for, which means in terms of what consumers find it worth while to pay for it. Hence, ultimately the valuation of the addition to output is the consumers' valuation of it. Therefore a shift from the firm where the value of marginal product is lower to the one where it is higher clearly must increase the value of total output. Also, it will pay to carry out such shifts from one firm to another until a point is reached at which the value of the marginal product is the same in the two firms. But the value of the marginal product of the factor will be the same in different firms in the case in which for each firm the value of the marginal product is equal to the competitive price of the factor, i.e., in equilibrium. Hence this is a situation consistent with a desirable allocation of the factors of production.

5

Monopoly and Monopolistic Competition

THE MONOPOLISTIC SELLER

We usually conceive of a monopolist as a man (or firm) who has complete control of the sale of a product in a given market. In other words, there is a complete absence of competition in the seller's market. However, the degree to which a so-called monopolist has monopoly power depends on the extent to which there are substitutes for his product. There are some substitutes for the vast majority of products, although the degree of substitutability is in some cases very close while in others it is quite remote. Where the substitutes are close, and many in number, the situation approaches that of competition. Where there are no substitutes we have pure monopoly, but such situations are rare indeed. However, for analytical purposes it is useful to consider the pure cases unencumbered by complications. Later in this chapter we consider the other extreme of many substitutes, the case of monopolistic competition. But first we must examine the pure monopoly case.

If a monopoly is to be maintained, then entry into the industry must not occur. Why it does not occur, or by what means it is prevented, we shall not examine at present. We assume that there is only a single seller, who acts as though no other sellers will or can enter the industry, and we examine the logic of his output and price decisions under these circumstances.

As before we suppose that our firm seeks to maximize profits. To simplify matters we assume that there is competition in the factor markets, and hence the firm in question is a price taker when it purchases *inputs*. Now maximizing profits for the monopolist, as for any other firm, implies the necessity of trying to minimize costs for any given output. Thus, the possible cost-output relations are the same for the monopolist as for the competitive firm. The main difference lies in the treatment of the demand side of the picture.

The basic power of the monopolistic seller that is not possessed by the competitive seller is that of setting the price. Monopolists are not price takers. But buyers will, of course, be influenced in their decisions by the price charged by the monopolist. The influence of price on buyers' decisions is shown, of course, by the demand curve. As usual, we assume that the demand curve is of the normal kind, i.e., it is negatively inclined. Hence, the higher the price that the monopolist sets, the lower the quantity demanded. His problem is to pick that price and output that will maximize his profits.

The logic behind the choice of the best price-output decision for a monopolist will be developed with the aid of Figure 12. The *AC* and *MC* curves in the figure are the usual average total unit cost curves and marginal cost curves for the firm. They are the same as those discussed in previous chapters. In the monopoly case we must first focus our attention on the *AR* (average revenue) and *MR* (marginal revenue) curves. The *AR* curve is the demand curve for the commodity. Since price is the same as the average revenue for the commodity, a point on the *AR* curve denotes the average revenue for the quantity in question. To every average curve there is a corresponding marginal curve, and in the figure the *MR* curve denotes the marginal revenue curve appropriate for the average revenue curve in question. The ordinate value of a point on the *MR*

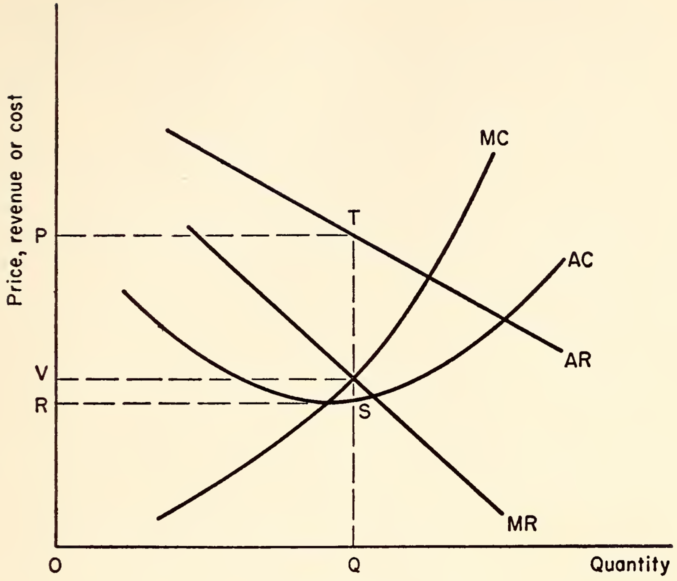


Figure 12.

curve indicates the additional revenue to be obtained by selling one more unit of the commodity. The negative slope of the marginal revenue curve (negatively inclined and below *AR*) represents the fact that in order for the monopolist to sell more of the commodity he has to lower his price on *all* units sold; hence additional revenue must fall, and it must fall to a greater extent than the average revenue. The average revenue falls only to the extent of the price reduction, but the marginal revenue reflects not only the lower price on the marginal unit sold but also the additional loss in revenue due to the fact that a lower price is received for all units sold.

As before, the optimal position is determined at the point of equality of the two marginal valuations. Thus the intersection of the marginal revenue and marginal cost curves determines the optimal quantity to be produced, the quantity denoted by *OQ* in the figure. The associated price *OP* is determined by the height of the point on the average revenue curve (or demand curve) opposite the optimal quantity *OQ*. We note that the profit-maximizing choice requires that two things be selected simultaneously: the optimal price *OP* and the associated best quantity *OQ*. This contrasts sharply with the competitive case, where the price was taken for granted and only the quantity had to be determined.

Now, the distinction between short run and long run is the same for the monopoly case as it is in the competitive situation. We need not consider the short-run and long-run situations separately and in detail here since the general decision problem and logic of the situation are pretty much the same in all cases. In this connection we can consider two alternate interpretations of Figure 12. Suppose that the analysis we have just carried out was the short-run analysis; all the curves we considered were those pertinent to the short run. We could easily reinterpret these curves in terms of the long-run situation. Suppose that the curve labeled *AC* is the long-run average cost curve and the one labeled *MC* the long-run marginal cost curve. Long-run price and output will be determined the same way as before. However, the optimal quantity *OQ* denotes now also the scale of operations. In other words, the optimal size plant will be one associated with the quantity *OQ*. As the figure is drawn we illustrate here the increasing cost portion.

In both short run and long run the outputs were equilibrium outputs. We note that in the long-run monopolistic equilibrium case there can be, and usually are, positive profits. The magnitude of the monopoly profits is shown by the rectangle *RPTS*. This is simply the profits per unit of *TS* multiplied by the number of units sold. Since we assumed at the outset that no entry can take place, we see that this situation persists in the long run. The rectangle *RPTS* represents the maximum profits possible, and there is nothing in the internal logic of the situation to cause the firm to make any price and output decisions other than those that lead to profits of *RPTS*.

Figure 13 illustrates the situation in the case of decreasing costs. In general the elements of the problem are practically the same as in the case of increasing costs. The solution, as before, is determined by the intersection of the marginal cost and marginal revenue curves. The optimal quantity is once again *OQ*, and it is associated on the *AR* curve with the optimal price *OP*. Once again we observe

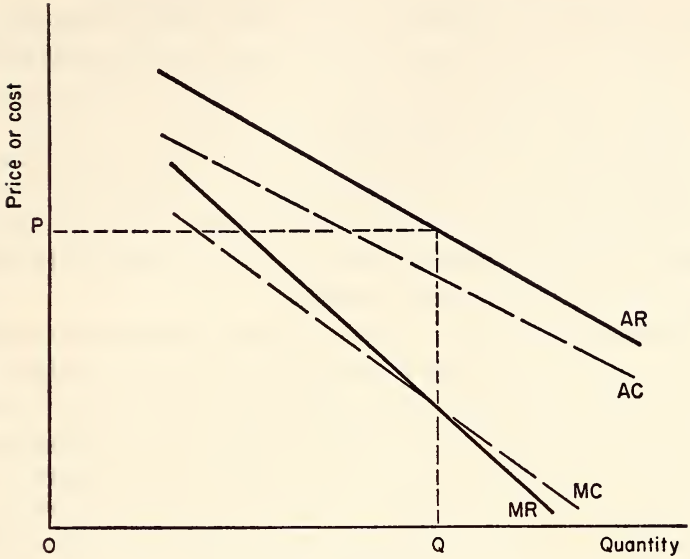


Figure 13.

the possibility of an equilibrium consistent with the persistence of monopoly profits.

Now, in the monopoly case there is no such thing as a supply curve for the commodity in the sense in which this concept was used in connection with the competitive firm or the competitive industry. While it is true that there is a unique supply quantity for the monopolist, given the demand curve of the commodity, there is no way of associating a quantity supplied to any given price. Since the monopolist has to determine both price and output simultaneously it does not make sense to consider a relationship in which we posit the price as an independent variable.

SOME PROPERTIES OF MONOPOLISTIC PRICE AND OUTPUT DECISION

We consider now some special aspects of the monopoly case. We notice at the outset that the monopolist does determine the price of the commodity. Furthermore, the price will, in general, be greater than the cost of production per unit. Also, the monopolistic firm need not produce at the lowest point on its long-run average cost curve, even where that curve is U shaped. It may pay the monopolist to have a firm size that is in some cases larger than the most efficient size from the viewpoint of the economy as a whole, and in other cases smaller than the most efficient size from the same viewpoint. Obviously there is no reason for the monopolistic firm to change its size at the expense of profits.

We assumed that the monopolist seeks to maximize his profits and hence to minimize cost of production for any given output. But there is really nothing that forces him to do so. He may be interested in a more relaxed way of life and will not strain to the utmost to produce at lowest cost. Thus the monopolist may produce whatever output he does produce in an inefficient manner as long as the demand price remains above his average cost. Within limits there is nothing to force him to produce efficiently. This is not the case with a firm in a competitive industry. In the long run the competitive firm is forced to minimize costs if it is interested in survival, as long as other firms try to minimize costs. Where competition exists the cost minimizer can undersell the non-minimizer and force him out of business.

The economists' criterion of economic efficiency is from the point of view of the consumer. Are the product mix and prices such that consumers get what they are willing to pay for? If the consumer is willing to pay x dollars for something, and the cost of shifting resources from other uses to producing the commodity in question is less than x dollars, then the existing allocation of resources is presumed to be inefficient from the consumers' viewpoint. It is in this sense that a profit-maximizing monopoly is usually presumed to be inefficient. In this connection let us look at Figure 12 again. Consumers are clearly willing to pay a price somewhat above *OV* for more of the product. The cost of additional resources is shown by the marginal cost curve, and they are no more than *OV*. Yet the monopoly will not supply any more output except at a price above *OP*. Hence resources are not utilized in accordance with consumers' desires under monopoly, as is evident from the fact that the monopoly price is above marginal cost.

MONOPSONY

The case of a single buyer facing many sellers can be analyzed in a similar fashion. We examine a typical monopsonistic situation with the aid of Figure 14. The main point is that the monopsonist can, to some extent, determine the price that he pays for the factor purchased. Suppose that the factor involved is labor. We have already indicated that the value of the marginal product of labor determines the number of units of labor that will be demanded. Thus the demand curve *D* in Figure 14 is also a value of marginal product curve. Now the more labor the monopsonist wants to buy, the more he has to pay per unit in order to attract it away from other uses. Hence, the supply curve is shown as positively inclined. The supply curve is simultaneously an average cost curve, i.e., it

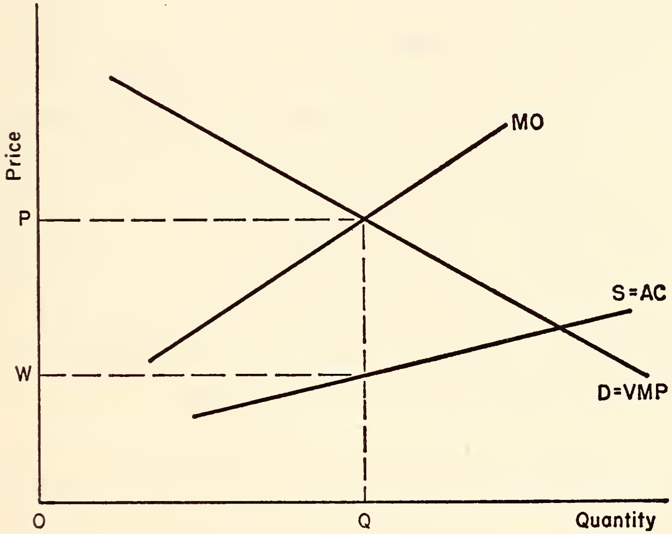


Figure 14.

indicates the price per unit of the factor. As usual there is a marginal curve to every average curve. The marginal curve that relates the alternate marginal outlays to alternate quantities of labor is denoted by MO in the figure.

The optimum quantity purchased is determined at the intersection of the marginal outlay curve and the value of marginal product curve. The logic is, of course, the same as before. As long as the additional value of output created by one more worker is less than the additional outlay necessary to hire that worker, it is desirable to employ him, and to continue to do so as long as this inequality persists. The optimal wage from the monopsonist's viewpoint is determined on the supply curve and is *OW*.

Once again we see that this involves an allocation contrary to consumers' preferences. Since the marginal value of the product created is greater than the wage, consumers' preferences will dictate that more laborers be in this industry than the monopsonist finds it profitable to hire. Under competition this would never arise, for the supply curve of any factor would be perfectly elastic and would be denoted by a straight line parallel to the *x*-axis in the figure. That is, under perfect competition among buyers of labor the marginal outlay and the wage would always be the same.

THE MULTIPLANT FIRM AND THE PERFECT CARTEL

Another example of monopoly decision-making is that of a decentralized multiplant firm. We suppose that the firm has a number of plants all of which produce the same commodity. The firm is a monopoly; it is the only supplier in the industry. Therefore, the demand curve for the industry is the average revenue curve for the firm. To make the problem somewhat more interesting we assume that the cost curves are different in the various plants. The problem faced by the monopolist is how to make three simultaneous decisions. At what price should he sell the output? What should be the quantity of output? And how should the production of the output be distributed among the different plants?

We can divide the problem into two parts: The first is the overall decision problem of determining the output for the firm as a whole, and the second is the problem of distributing the output among the different plants. Consider the first part of the problem. The demand curve for the industry is the average revenue curve for the firm, and from it we can derive the associated marginal revenue curve. Thus we have the usual *AR* and *MR* curves for the firm. The cost curves of the firm are the cost curves of the individual plants that make up the firm. Thus our immediate task is to aggregate the plant cost curves to obtain the cost curves of the firm. We can aggregate the marginal cost curves by adding them horizontally in the same way as we might add the marginal cost curves of competitive firms to determine the industry supply function. This situation is illustrated in Figure 15, where we assume that there are only two plants. For each marginal cost value we simply add the appropriate output from plant A's marginal cost curve and from plant B's marginal cost curve to obtain a point on the firm marginal cost curve. If it were necessary we could also obtain the firm's average variable cost curve in a similar manner. However, the firm marginal cost curve is all that is necessary, although we must be careful to include only that portion of a plant's *MC* curve that is above average variable cost.

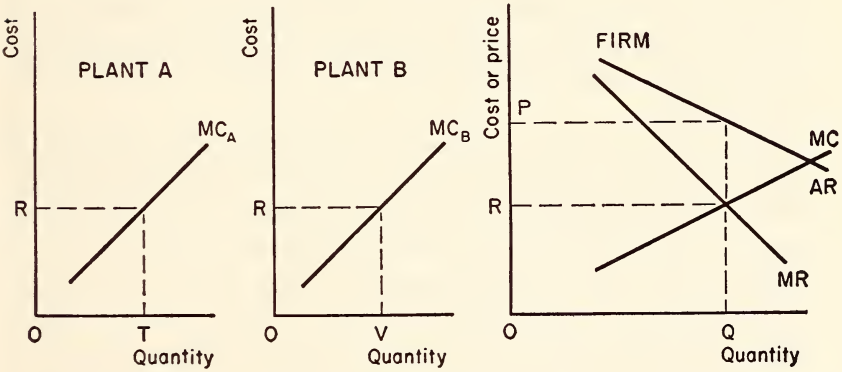


Figure 15.

The firm decision is exactly the same as in the usual monopoly case. The intersection of the marginal revenue and the marginal cost curves determines the quantity to be produced, and as usual the ordinate on the average revenue curve determines the price. As shown in the figure the optimum quantity is *OQ*, and the optimum price is *OP*.

Now the value *OR* shown in the figure is the element that enables us to solve part two of our problem. *OR* is the marginal revenue to the firm for the quantity *OQ*. To distribute the quantity *OQ* between the two plants we have to ask each plant to produce that quantity at which its marginal cost is equal to *OR*. This yields an output of *OT* for plant A and *OV* for plant B. Clearly *OT* plus *OV* must be equal to *OQ*, since we obtained the quantity *OQ* on the firm marginal cost curve by adding the quantities on the plant marginal cost curves associated with the marginal cost equal to *OR*.

The same procedure could be used for a cartel. If the cartel is a perfect cartel then the procedure just outlined would maximize cartel profits.

The profit distribution arrangement among members of the cartel must be entirely separate from the output of any member firm. These should be treated as separate questions. For example, there may be a firm whose plant's average variable cost curve is (for all outputs) above the marginal revenue of the cartel at the cartel's optimal output, even though the cartel monopoly price is higher than the average total unit costs of the plant in question at some of its outputs. The attempt to maximize cartel profits would dictate that the plant in question should not produce anything. Only in this way would the cartel output be efficiently produced. However, the relatively high cost firm can claim a share in the cartel profits. Otherwise it would not join the cartel but would produce some output and earn profits on its own, as long as it could produce and sell at a price equal to or below the cartel average revenue, although at the same time its marginal cost would be above the cartel marginal revenue. Therefore, if the cartel is to be a complete monopoly it must induce the inefficient firm to enter the cartel, and to do so it must offer it a share in cartel profits.

MONOPOLISTIC COMPETITION

A set of situations in between pure monopoly and pure competition are those usually designated under the rubric of monopolistic competition. In monopolistic competition we visualize a market in which there are a great many relatively small firms. However, there is a monopolistic element in the sense that each firm has some "hold" on a portion of the market. The "hold" may exist because of some special quality associated with the commodity that the customer believes does not exist to the same degree for substitutes sold by other firms. As a consequence each firm has a slight monopolistic advantage since no other firm sells a commodity quite like it, but the other commodities in the group are all very close substitutes.

This sort of situation may arise even where from a purely physical or technical viewpoint the commodities are exactly alike, but because of the imperfect information available to consumers they appear to them to be different. For example, trade marks, differences in packaging, and so on often create in the minds of consumers the belief that there are differences in the characteristics of commodities that are in fact essentially the same. As a consequence we can say that each producer under such circumstances enjoys a monopolistic advantage to a small degree. It is of special importance to note that a reduction in the price of a close substitute will not mean that a firm will lose all of its sales to the competitor, as is the case under pure competition. Therefore, for each firm under monopolistic competition the average revenue curve is not infinitely elastic (a horizontal line as in the case of competition) but is rather a somewhat negatively sloped curve, somewhat comparable to that of pure monopoly.

Since the commodities for the various competitors that we have in mind are not perfect substitutes for each other we cannot call our group of competitors an industry in the strict sense of the word. Following Chamberlin,[1] we simply call the set of firms offering close substitutes the monopolistically competitive group. One of our problems is eventually to determine the conditions of group equilibrium. But before we can do so we must consider the decision problem of the individual firm. The basic decision problem of the firm in monopolistic competition is precisely the same as that under monopoly. We can, therefore, treat this matter with dispatch. The firm has to determine an optimal quantity and the optimal price associated with that quantity. It will pick the price-quantity combination, given the relevant costs and average revenues, that will maximize profits. Once again the firm will pick that combination associated with the intersection of marginal revenue and marginal cost.

---

[1] E. H. Chamberlin, *The Theory of Monopolistic Competition,* 7th ed., Harvard University Press, Cambridge, 1956.

There is a significant additional element or set of elements to be considered here which we may conveniently put under the rubric of selling costs. The firm may vary not only price and output but also the amounts of selling costs (advertising expenditures, packaging costs, etc.) that it wishes to put into the commodity. Such expenditures will, in most cases, influence the price at which a given output can be sold. That is to say, sales expenditures will, in part, determine the shape of the demand curve facing the firm. Hence, we end up with a three-element decision problem: that of output, price, and selling cost determination.

By selling cost we have in mind more than just advertising. Such elements as the packaging of the commodity and other expenditures that differentiate the commodity from its competitors, alter the looks of the commodity, or in some way give a different impression of the commodity to the consumer are to be included. To a considerable extent the firm will want to differentiate its product from the products of all other sellers of substitutes so that it can keep its slight monopoly hold on at least a portion of the consumers. Thus any costs that can in some way help to increase demand, or that for a given quantity can increase the demand price at which the commodity can be sold, should be included in selling costs.

We therefore distinguish carefully between the manufacturing costs of the commodity as such and the selling costs. Of course, the final cost to be considered in making the price-output decision is the sum of the manufacturing costs and the selling costs.

Selling costs will have three effects: (1) they will shift the demand curve to the right; (2) they are likely to change the shape and slope of the demand curve and make it somewhat less elastic than it otherwise would be; (3) they will, of course, add to the cost of producing the commodity for the market.

We now consider the decision problem in two steps. First, for each alternate output let us determine what the optimal selling costs are for that output. The effect of selling costs, we recall, is to enable us to sell the output at a higher price, but at the same time this necessitates an increase in the cost per unit of output. As long as a dollar's worth of selling cost will add more than one dollar to net revenue, we will want to make that selling expenditure. The logic involved in determining the optimal selling expenditures for

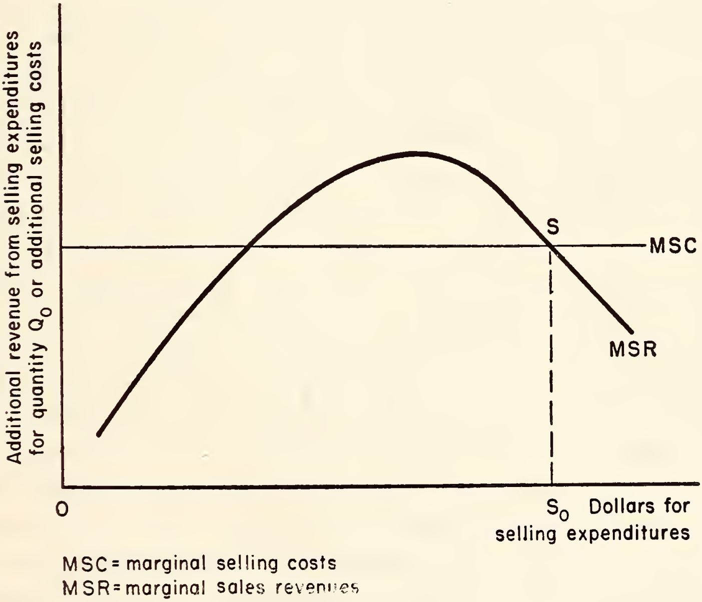


Figure 16.

a given quantity of output Q_0 is illustrated in Figure 16. The ordinate in the figure indicates the net revenue for that quantity. The abscissa denotes alternate amounts of selling costs, that is, different dollars' worth of selling costs. The curve *MSR* indicates the selling cost-marginal revenue relationship. For each amount of selling cost it indicates the addition to revenue that will be obtained. As drawn in Figure 16, in the early part of the curve the effect of increasing selling costs is to increase the additional revenue obtained more than proportionately. This is a common experience. For instance, as advertising expenditures increase, up to some point, there is a maximum effect; beyond that point there will be diminishing returns.

The optimal selling cost for the given output Q_0 is that amount associated with the intersection of the *MSC* (marginal selling costs) and the *MSR* (marginal sales revenue) curves—that is, the point at which an additional dollar's expenditure on sales effort will *not* yield more than an additional dollar of sales revenue. This intersection is shown by the points *S*, and the optimal amount of selling cost is indicated by OS_0. The curve *MSC* in the figure is, of course, a marginal (as well as an average) selling cost curve. The total revenue obtained out of this selling cost is the area under the curve *MSR*, and the total expenditure is the area under the line *MSC*. If we divide these amounts by the quantity of output Q_0 we obtain the addition to the price at which we can sell the quantity in question, and the addition in selling costs per unit involved in the optimal selling cost. Thus Figure 16 enables us to determine for a given quantity of output a point on a *new* average revenue curve.[2] It also enables us to determine the amount of selling costs per unit of output. Adding this last amount to the manufacturing cost for the relevant quantity we get a point on the average total unit cost curve. In this way we obtain a new point on a demand curve and a new point on the average total unit cost curve.

By following the same procedure for all of the alternate quantities of output we obtain an average revenue curve, and an average total unit cost points inclusive of selling costs; i.e., we obtain the average revenue and cost curves necessary to make our final decision. We then proceed in the same way as under the monopoly case. As usual from our average revenue curve we can obtain the marginal revenue curve, and from our average cost curve we can determine the marginal cost curve. Below the intersection of the two curves we may read off the optimal quantity of output, and the associated point on the average revenue curve yields the optimal price.

---

[2] It is to be observed that this is not the usual type of average revenue curve. Each point on this curve implies an optimum amount of selling costs for the quantity of output in question.

GROUP EQUILIBRIUM

Suppose, to begin with, that all the firms in the group earn profits. To retain the competitive aspect of monopolistic competition we assume free entry. New firms enter on more or less the same terms as the existing firms. Also, monopolistic profits induce firms to enter the area. Now, the effect of new entrants is to reduce the demand for any particular firm. The new entrant creates substitutes for the commodity of every firm in the group. As a result some customers may prefer the product of the new firm at its price as against the product of any existing firm. In sum, we visualize that the effect of entry is to shift the demand curve of all existing firms to the left. This shift, of course, will also reduce the profits of all the existing firms. Indeed, the process of new firms entering, of the curves shifting to the left for all firms, and of profits falling will continue as long as there are any profits being made by members of the monopolistically competitive group.

An equilibrium is struck when the profits of all of the firms are zero, so that there is no inducement for any potential firm to enter or for any existing firm to leave the group. This is illustrated in Figure 17. The monopolistically competitive demand curve (*D*) is tangent to the average total unit cost curve (*AC*). Only under the condition of tangency will there be no monopolistic profits for the typical firm in the group.

Now we consider some of the properties of group equilibrium. First, the price, as under competitive equilibrium, is equal to the cost of production. As usual, of course, we include in our cost calculations the opportunity cost of management and of entrepre-

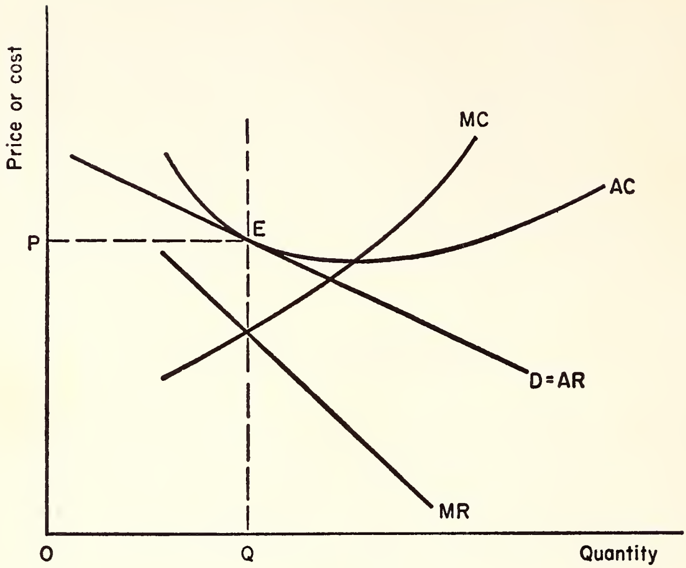


Figure 17.

neurial effort. From this point of view the group equilibrium appears similar to the competitive industry equilibrium. However, there are elements in the monopolistically competitive situation that are in sharp contrast to the competitive equilibrium case. For example, the average total unit cost in this case is above what it would be under pure competition, since the total average unit cost includes selling cost. And presumably selling costs exist because the firms are operating in an imperfectly informed market, and they are costs for activities that go beyond the simple provision of accurate market information.

In addition, we observe that the firm's output is not at the lowest point of its average total unit cost curve, partly because the demand curve is negatively inclined under monopolistic competition and therefore the point of tangency between demand and average cost cannot possibly be at the lowest point on a U-shaped average total unit cost curve. Hence, from this viewpoint also the monopolistically competitive equilibrium would appear to be a less efficient state of affairs than the purely competitive one.

6

The Oligopoly Problem

ONE of the least satisfactory portions of modern economic theory is that segment which deals with oligopoly. The reason for this is the intrinsic difficulty of the problem. This is an area in which a long history and a huge literature have developed. It will not be possible, of course, in the pages that follow, to review all of that literature or to even consider a great many of the solutions that have been suggested. All we can attempt to do at this stage is to indicate the nature of the problem (this in turn will suggest why it is difficult to solve it in complete generality) and outline a few rather simple models that attempt to cope with some elements of the problem.[1] We will have more to say on oligopoly in the last chapter of this book.

The central element of the traditional approach to the oligopoly problem is decision-making in the face of "circular interdependence." This can, perhaps, be easily perceived in the case of duopoly. Here we consider two sellers that sell the same kind of commodity in the same market. Various strategies are open to them—in general a larger number than is possible in either pure competition or pure monopoly, or even in monopolistic competition. The central difficulty lies in the fact that each seller cannot take his environment as given. Firm A does not know what firm B will do, nor does firm B know what A will do. Yet the action of each will affect the other, and the outcome. As a consequence no one seller controls all of the variables of the situation in such a way as to be able to predict the outcome of any given action on his part. Therefore it is impossible for any one seller to make an optimal decision in the usual sense of the word, i.e., a decision that he is assured will lead to the best outcome from his viewpoint.

---

[1] For an excellent discussion of what might be called the classical oligopoly problem see William Fellner, *Competition Among the Few,* Alfred Knopf, New York, 1950.

Consider some of the alternatives for firm A. First, firm A can set his price without reference to the price set by firm B. This, of course, would be rather risky for should firm B set a lower price, firm A would lose all of his customers to B. Therefore, while this is a possible strategy it is hardly a wise one in the long run. Second, firm A could attempt to be a price follower with respect to B, and accept whatever profits result from this tactic. The wisdom of this strategy is that there is some safety in it. Firm A will not lose all of his customers as a result of this procedure. Third, firm A could attempt to be a leader. Mainly, he may try to set the price in the hope that firm B will follow it. Should it turn out that firm B does choose to be a follower, then, of course, firm A is in a rather advantageous position. He can set that price which will maximize his profits. Fourth, firm A could decide to engage in a price war and force firm B out of the industry. Should firm A win such a war, of course, he would be a monopolist and could make even higher profits than he would with any other strategy. However, he takes the risk of being the loser, and in addition there is the cost of a price war to consider even if he turns out to be the winner. A fifth strategy is for firm A to be a price follower but to attempt to compete in other areas. For example, firm A may be more or less a price follower but try to give superior services to customers or compete with respect to quality differentiation, advertising, packaging, auxiliary services, and so on. Now the significant element in all of these potential strategies is that there is no way for firm A to choose a strategy that is best for him in the abstract. It all depends on what he believes firm B's strategy will be, and in detail what firm B's countermoves will be to his initial moves. As a consequence, this is not a simple maximization problem as we usually conceive it in economic analysis. All of this will come out more clearly when we consider some specific examples.

SOME POSSIBLE EQUILIBRIUM POSITIONS

Suppose that our two duopolists are similar in every respect. That is, both firms are exactly the same size, both cost curves are identical, and each shares one-half of the market. In other words, we assume that when both sellers charge the same price each gets half the trade at that price. The demand curve for each firm is the same. Now, we can visualize that for each firm there is also a marginal revenue curve and the usual marginal cost curve. Therefore, there is a price which will maximize the profits of both firms, and of each firm separately, if each charges the same price. But this profit-maximizing price—let us call it P_0—may or may not be the equilibrium price. Indeed, it may turn out that there are many equilibrium prices, or that there is no equilibrium price, depending on how the two firms react to each other's behavior, and on how each one believes the other will react to his behavior.

Each firm may set his price without taking cognizance of the other fellow's price. In that case, they would immediately set their price at the optimal one P_0, and if they continued to behave this way—that is, if they continued to ignore the other's behavior—the situation, of course, would be stable.

But it is quite possible that firm B may attempt to get additional profits by capturing some of the market from firm A. In that case B may set a somewhat lower price than A in the hope or belief that firm A will not drop his price accordingly. If A does drop his price to the same level as B's, then, of course, both firms are worse off. Now, there is no clear-cut way to determine how the two duopolists will behave from here on in. Consider some of the possibilities. B may decide that his initial move was a bad one and he may attempt to raise his price back to his optimal price. If A follows him by raising his price, we again have a situation in which the initial optimal price P_0 is a stable equilibrium price. A second possibility is that when B raises his price A does not follow suit. It may be to A's advantage to remain at a lower price than B and capture some of B's market. In that case B may, after he observes that A does not follow him in raising his price, reduce his price again to the lower level. Furthermore, neither of the two may then raise their prices in fear that the other will not follow. This is perfectly reasonable behavior and as a consequence the lower price is now an equilibrium price. Quite clearly, should there be another move on the part of either duopolist toward a lower price it will quite possibly be met with an additional lowering of price by the second duopolist, when this new price will become an equilibrium price. In other words, the situation is asymmetric. There is a strong inducement for each duopolist to follow with respect to the lowering of price by one or the other, but there is no inducement for either of them to do so in the case of a rise in price by one of them. We therefore see that there is a whole series of prices which may or may not be equilibrium prices depending on the tactics, the psychology, and the expectations of each of our duopolists.

In the above situation there is no unique solution unless we assume something specific about the knowledge that each duopolist *may have* about the other's potential behavior. This leads us to the subject of reaction curves.

REACTION CURVES

A scheme sometimes used to illustrate the nature of the duopoly problem, and one that is sometimes put forward as a "solution," is that of reaction curves. These are illustrated in Figure 18.

If we are given the price of firm A we might assume that firm B will find it convenient to react in a certain way. Where the two commodities are precisely identical, and there are infinite elasticities of demand, we know that both firms will want to sell the commodity

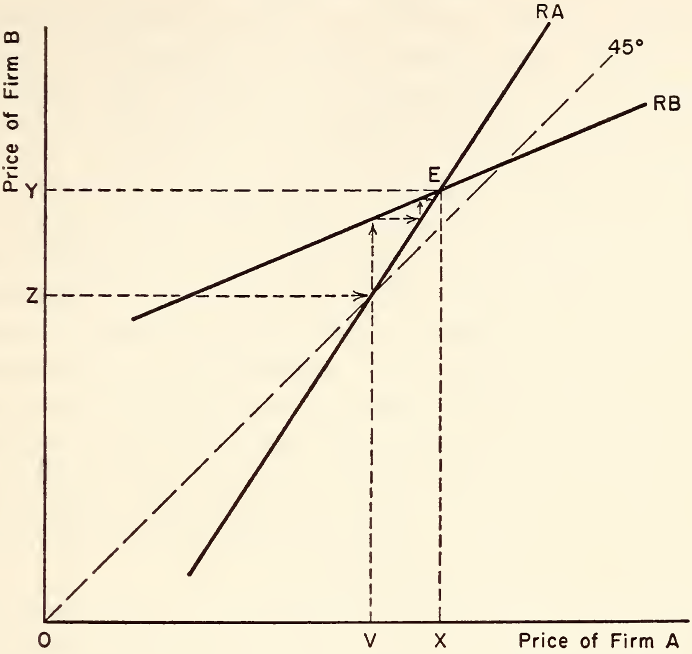


Figure 18.

at the same price. Under that situation if one firm should lower the price the other will try to follow immediately and vice versa. Thus, in the case of exact substitute commodities characterized by very rapid reactions by consumers to price differences, there will be a large series of prices that could be equilibrium prices. The situation is somewhat simplified if we consider the case of a somewhat imperfect market, that is, one in which firm A does not lose all of his customers to firm B simply because he sells his commodity at a slightly higher price. This case is often illustrated by the use of reaction curves.

The curve represented by *RA* in Figure 18 is the reaction curve for firm A. What it says is briefly as follows: For every price of B there exists a price of A that firm A will want to have, given the price set by firm B. For example, at very low prices set by firm B, A will not mind setting a price above that of B. Although in that case firm A will lose some of the market, it will gain by the fact that it sells each unit of the commodity at a higher price. Thus, at low prices firm A's reaction curve will present higher prices than those set by B. On the other hand, at very high prices A's behavior will be in the opposite direction to B's. A will prefer to set a price below B's, since at high prices consumers are presumably very sensitive about price differences between different firms. As a consequence the gain from a larger share of the market will be more important to firm A than the slight loss due to the fact that he sells each unit of the commodity at a lower price. Therefore, if we draw a 45° line through Figure 18 we will find that a portion of A's reaction curve will be to the right of that line and another portion to the left. Now we may assume that firm B will react in a somewhat similar, though not exactly the same, fashion. (We could assume that B will react in exactly the same way as firm A but we need not make so stringent an assumption.) The curve denoted by *RB* in the figure is firm B's reaction curve.

Where the two curves intersect we have an equilibrium position. Namely, we have a set of prices set by firms A and B which will be stable. For example, the price *OY* set by firm B will induce firm A to set a price of *OX*, which in turn will induce firm B to maintain his price at *OY*, and in turn this will induce firm A to maintain his price at *OX*, and so on. Clearly, these two sets of prices for the commodity are stable. Thus, this is a duopoly situation in which there would appear to be an equilibrium solution.

Furthermore, if we assume that the two reaction curves are themselves stable, that is, if they are fairly permanent for all such possible prices set by one or the other, then we can see that the equilibrium set of prices is also a stable equilibrium. For if we presume a deviation from that equilibrium, we can see that the path followed by the two firms will be that indicated by the set of arrows in the figure. This path of change is always toward the equilibrium set of prices denoted by the point *E* in Figure 18.

The main difficulty with reaction curves is that outside of the strict equilibrium case they appear to be based on rather odd assumptions about expectations. That is to say, it is assumed that each firm expects the other firm will maintain his price, or behave as though he expects the other firm to maintain his price. If firm B sets a price of *OZ* then firm A will set a price of *OV*. But this will induce firm B to set a different price, a somewhat higher price, and firm A will react to that price, and so on. As a consequence, every time we have a set of moves each firm must discover that the other firm's position is not a stable one. But we assume that each firm reacts as though the other firm's position is stable. In other words, we assume that each firm in making his decision will not take into account what the other firm's reaction to his decision will be. If we assumed otherwise, of course, the situation would become infinitely more complicated and could not be illustrated by the sort of reaction curves that we employed. However, this does not affect the equilibrium situation, the one denoted by the point *E*. Thus, we can still say that the point *E* represents an equilibrium position if somehow it is struck. What we can say about the other cases is that they do *not* represent equilibrium situations, and very little more. Whether or not points other than *E* generate a path of mutual reactions that lead to a return to point *E* will depend, of course, on each firm's expectations as to how the other firm will react to this move or countermove. To answer this question in detail would require that we develop a theory telling us how any sequence of moves alters the expectations of the firm in question.

MARKET STABILITY, PRICE LEADERS, AND PRICE FOLLOWERS

Although we have indicated previously some reasons why oligopoly markets might be very unstable, we find, in fact, that a large number of such markets do have a high degree of stability. One economist has compared the oligopoly market situation to the issues of war and peace.[2] But periods of peace are likely to be longer than periods of war. Similarly, the oligopoly situations may sometimes break out into price wars and wars of survival. Nevertheless, in most markets the period of truce is likely to outweigh that of war. By their very nature and costliness price wars are usually temporary, whereas the periods of seeming stability in the market structure are inclined to last longer and be the ones that are observed.

---

[2] Cf. K. Rothschild, "Price Theory and Oligopoly," *Economic Journal*, September, 1947.

We can easily envisage a situation in which the various firms in an oligopolistic industry do not find it worth while to engage in price wars. The gain from a price war is either to force some other firms into bankruptcy or to get a larger market share. In essence, then, a price war is a war for market shares. However, such wars are generally quite costly and it is not completely determined which firm will come out on top. As a consequence, in many situations the expected gains in terms of market share do not compensate for the possible losses involved. Furthermore, the loss might be of an extreme kind. If the firm risks bankruptcy by entering into a price war, it might very well feel that such a risk is too large. To avoid the struggle for market shares through price competition the firms have to find some means of getting along without the risk of engaging in such struggles. One obvious means of doing so is collusion; they may get together and agree on a market price. However, this may be quite difficult. There are often legal or other barriers to outright formal collusion. But when all firms in an industry seek a means to avoid a price fight they will usually find some means even if they cannot enter into a formal agreement. This very often takes the form of price leadership. The other firms in an industry will recognize some firm as the price leader and will set their prices in accordance with the price set by him. When the price leader raises his price they in turn will raise theirs; similarly when the price leader lowers his price they in turn will lower theirs. Who becomes the price leader is usually determined, in part, by the history of the industry. Clearly, though, if a pattern of price leadership is established, and if each firm feels that a struggle for market shares is not worth the risk, given the potential gains, then this will be a fairly stable industry.

In some cases the price leader will turn out to be the largest or most powerful company in the industry. It may be the company that in any event can enforce its will on the others, and that could outlast the others in a price war. However, this need not always be true. As we have already suggested, which firm becomes the price leader will depend on the history of the industry in question. A relatively small firm may be highly respected and efficient, and some incident in the history of the industry may move it into the position of price leader. In other words, at some point some of the companies may have decided to follow the smaller company in its price policies, and the largest company in the industry may have found it worth while not to resist this trend. Once such a practice and tradition is established, the small company will remain as the price leader in that industry. Not only is the weight of tradition likely to be exceedingly important in such situations, but in addition no company in an oligopolistic industry may find it worth while to challenge the price leader for its position since doing so involves the risk of a price war with *all* firms in the industry. This is not to suggest that price leaders are never challenged; it is only to suggest that the fear of a price war may deter anyone from upsetting any pattern or scheme of avoiding price competition that develops.

Now, the practice of price following will not, in the usual case, lead to a price that is best for *every* firm, i.e., it will not be the price that will maximize profits for every firm. However, the oligopoly situation is often one in which getting along with other firms and not risking a price war are more important than trying to maximize profits. Clearly, survival in an industry is more important to a firm than merely *attempting* to maximize profits regardless of the risk. Indeed, we have already indicated that, strictly speaking, profit maximization is not meaningful in this context.

ELASTICITY OF DEMAND AND OLIGOPOLY

An earlier chapter mentioned that the elasticity of demand may be viewed as a measure of customers' responsiveness to a change in price. Now, from the point of view of a firm in an oligopolistic industry, the way customers react to a change in price will depend not only on the magnitude of the price change but also on whether other firms follow suit or not. It has been suggested by some writers[3] that the elasticity of demand for a reduction in price is quite different from that for a rise in price. The important consideration is that the firm believes that the demand for its output is likely to be inelastic with respect to price reduction and elastic with respect to price increases. The firm feels that if it lowers its price, its competitors will follow suit, while if it raises it, the competitors will not raise their prices. A reduction in price by a firm will mean that it will get no increase in the market share, since other firms follow suit, but an increase in price will cause it to lose some of its market shares. Hence, the firm sees no point in lowering its price, but it dares not raise it unless the recognized price leader does so first. This mode of analysis has been used to explain presumed price stickiness in oligopolistic industries.

ENTRY-DETERRENT PRICING

In the previous sections we assumed that the number of firms in the industry is given at the outset. But oligopolists will in many cases have to worry not only about their existing competitors but also about potential competitors. Hence, they have to be concerned with the conditions of entry and if possible choose policies that will deter entry. Indeed some writers have suggested recently that entry deterrence is a significant determinant of oligopoly pricing.[4]

While it may be exceedingly difficult to say just how oligopoly prices are set, something may be said about the bounds within which they will be set during a period of truce, given that the oligopolists have some means of agreeing on prices or pricing procedures. Clearly, the lower bound in the long run is a price equal to the minimum cost of production. But the upper bound may be less than a profit-maximizing price for any firm in the industry if there is ease of entry. Obviously, one interest that all oligopolists have in common is to prevent or deter new firms from entering the industry. Firms would be deterred from entering (although strictly speaking they would not be prevented from doing so) if the prospective price after entry were below the long-run cost of production of the entering firm.

---

[3] See Paul M. Sweezy, "Demand Under Conditions of Oligopoly," *Journal of Political Economy*, August, 1939.

[4] See Franco Modigliani, "New Developments on the Oligopoly Front," *Journal of Political Economy*, June, 1958, for an exposition of the analysis of P. Sylos Labini and J. S. Bain on entry-deterrent pricing. The diagrammatical exposition is adapted from Professor Modigliani's article.

First, consider the case in which existing firms have a cost advantage over newly entering firms. Let us suppose that this is *x* percent above the minimum average total unit cost of production of the highest-minimum-cost firm in the industry. In that case the existing firms will deter entry as long as they set a price, say Px, which is slightly less than *x* percent above the minimum unit cost of the least efficient firm already in the industry. Under such circumstances new firms would be deterred from entering the industry, since they could not expect to earn long-run profits at a price of Px or less. There is surely no reason for the potential entrant to expect that the price will be above Px after it enters the industry.

A firm considering entering an industry is not primarily concerned with what the price is prior to entry. It must try to assess what the price is likely to be after entry. Hence, if the current price and related output conditions of the industry are such that the price after entry would lead new firms to lose money, then this would be an entry-deterring price.

Now, suppose that new firms could enter on the same terms, from a cost point of view, as those already in the industry. It has been argued that even in this case there are some circumstances under which the entry-deterrent price can be above the minimum cost per unit. Without attempting to defend the realism of the assumptions involved we present here, briefly, the essence of the argument.[5] The basic assumptions are as follows: The technology of the industry is such that the long-run average cost curve is L shaped, and the decreasing part of this curve is very steep. (See Figure 19A.) In other words, because of certain technical indivisibilities the minimum plant size (and firm size) must be fairly

---

[5] *Ibid.*, p. 217.

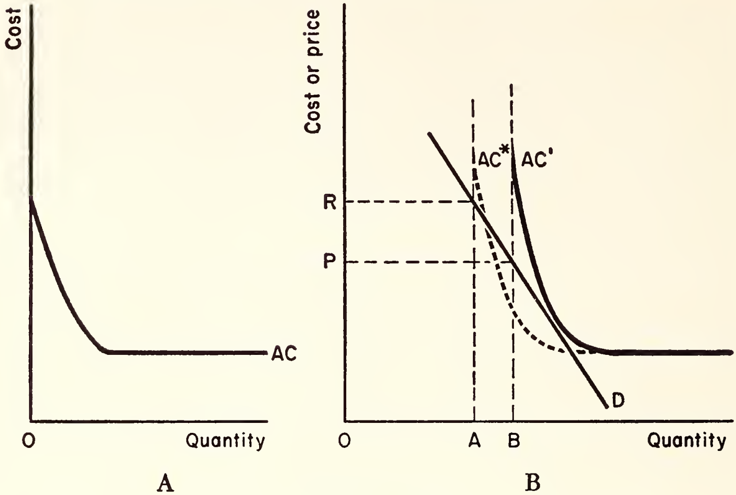


Figure 19.

large or else average costs of production are prohibitively high. Also, it is assumed that the minimum output at lowest cost for a single plant is a significant proportion of industry output. Thus there is no room for a great many plants in the industry. It is further assumed that "potential entrants behave as though . . . they expect existing firms to adopt the policy of maintaining output while reducing the price to the extent required to enforce such an output policy."[6]

Now, suppose that there are not many plants of the type illustrated in Figure 19A in the industry and that each plant has to produce, say, at least a million units in order to produce at minimal cost. Clearly, if a new firm enters and builds an additional plant, and if it produces at minimum cost, the output of the industry will expand considerably and the demand price will fall quite a bit. The problem of the existing firms is, in part, to determine a price and output policy that will deter the entry of a new firm and the building of a new plant. They may be able to find it if they currently produce an output sufficiently large so that the additional output of an additional plant would bring the demand price below the unit cost of the additional plant. The problem is to determine what such a minimum output would be since, given the minimum output, we can also determine the price on the industry demand curve at which that output could be sold. This is illustrated in Figure 19B. The curve *D* is the industry demand curve and the curve *AC′* is the long-run average cost curve of an additional firm with one plant. If the industry produces *OB,* the new firm's output will simply be added on to the industry output of the existing firms. We see from the figure that in this case there is no point at which the industry demand curve will be above unit cost for the new firm. On the other hand, if the industry produced only *OA* and the new firm's output was added to that, as shown in the figure by *AC**, then the demand price would be above the minimum cost per unit of the new firm, and, in this case, entry would not be deterred. We observe, therefore, that the price *OP* would be an entry-deterrent price while the price *OR* would not be an entry-deterrent price. If *OR* is the current profit-maximizing price for those in the industry, the members of the industry might be wise to charge the price *OP* rather than the short-run profit-maximizing price *OR,* which would invite entry.

---

[6] *Loc. cit.*

It will be noted that we used the phrase "entry-deterring" and not "entry-*preventing*" pricing. For no matter what the price and output policies of the existing firms, there is, in fact, nothing in the picture we have described to prevent a stubborn entrepreneur with sufficient financial resources from trying to force his way into the industry, selling below costs for a period, and if necessary engaging in a price war with the existing firms. If he survives the price war, then, of course, he will successfully have forced his way into the industry. But this must mean that some other firms either have been forced out of the industry during the price war or have reduced their output and share of the market to provide room for the new firm. Nevertheless, the sort of pricing scheme we have depicted may act as a deterrent to entry since potential entrants, if they are aware of the costs and market conditions in the industry, will be put on notice that they must fight their way into it.

Part II

Specialization

7

On the Meaning of Specialization

ALTHOUGH the concept of specialization is an old and hoary one in economic analysis, it is not always appreciated that there are some inherent problems of meaning connected with it. In this chapter we examine briefly why this is the case, while in the next chapter we review the main elements that determine the degree of specialization.

Many of us have read Adam Smith's analysis of the division of labor[1] and his description of specialization in a pin factory. Despite its beginnings, the idea of specialization is not treated formally in economic analysis, although it is often invoked in an *ad hoc* way to explain such phenomena as decreasing costs and increasing returns to scale. We often attribute economies in production to an increase in the degree of specialization. In many cases we believe that, up to a point, the more specialized a process (i.e., a way of producing a commodity, or of doing something) is, the more productive or efficient it is. But what is meant by a "more specialized process"?

---

[1] Adam Smith, *The Wealth of Nations,* Modern Library Edition, New York, chaps. I to III.

Specialization, including the division of labor, is also significant in underlying the questions of organization. Indeed, a change in the degree of specialization by which certain tasks are performed often creates organizational problems. Thus, the concept of specialization may be looked upon as a convenient transitional idea between microeconomic theory and organizational analysis. But if we are to use this notion, either explicitly or implicitly, then we should face the question, What do we mean by it? Hence this brief inquiry into the question of meaning.

There is a general feeling that we know what we mean by specialization, increased degree of specialization, increased opportunities for specialization, and so on. In many cases this intuitive notion is correct. Yet we shall see that the way we use these terms is in essence no different from the way we employ such concepts as rich, wealthy, poor, and so on. The essential point is that not in all cases is it possible to compare two production processes and say which of the two involves a greater degree of specialization. We shall see that (1) in some cases it is completely clear what is meant by an increase in specialization or an increase in the degree of specialization, (2) in other cases it is not obvious but by making certain assumptions we can give a meaningful interpretation to this concept, and (3) in still other cases it is impossible to compare degrees of specialization meaningfully.

COMMODITY SPECIALIZATION VERSUS ACTIVITY SPECIALIZATION

In economics the concept of specialization is usually employed with respect to commodities. For example, in international trade theory countries achieve the advantages of specialization by producing a narrower rather than a wider range of commodities. While it is true that in Adam Smith's example of a pin factory specialization in terms of productive activities is emphasized, it is nevertheless a fact that the commodity rather than the activity interpretation has pride of emphasis in contemporary economic analysis. Also, the tradition of analyzing production in terms of generalized inputs and outputs leaves out the possibility of giving any special consideration to productive activities. Since we shall emphasize in what follows the interpretation of specialization from the point of view of productive activities, it may be desirable to dwell briefly on the advantage of this approach as against the commodity approach.

Let us suppose for a moment that the nature and range of the commodity spectrum is given at the outset. Assume also that each commodity is unique and different in a discrete sense from every other commodity. In this case we may speak of a greater degree of specialization as being one in which a smaller rather than a larger range of commodities is produced. Within limits this is a fairly clear-cut concept. But some reflection should indicate that it by-passes a number of problems and distinctions. For each commodity there may be more than one way of producing it—indeed there usually are many. Some ways of producing the commodity may be said to be more specialized than others. Clearly, there would seem to be little reason not to recognize the second type of specialization in our analysis. On this ground the activity rather than the commodity approach appears to have the edge. But we need not ignore, or banish from consideration, specialization in terms of commodities. Our suggestion here is simply that activities are more basic units than commodities when we come to look at specialization within the context of the analysis of production, or in the context of what goes on *within* a firm or other type of organization.

THE SIMPLE DIRECT LABOR CASE

Is it meaningful to say that one production process is more specialized than another? By a production process we have in mind a collection of activities whose performance will produce a given commodity. Let us begin by considering a few simple cases and then proceed to more complex ones.

For our first case we make the following assumptions: The commodity has a definite physical form with a given unique set of characteristics and specifications. The production processes involve the transformation of a given raw material (steel, iron, etc.) from one form or shape into some other. We now consider two production processes, both of which could lead to the creation of the same commodity, and compare them with respect to the characteristic we call specialization.

At this stage we can define the notion of an activity in more or less natural terms. Since there is a raw material we can speak of the action of some of the other factors of production on the raw material. Of course, some of the factors of production will probably not act directly on the raw material but will be engaged in activities that in a sense may be said to facilitate the direct activities on the raw material. Thus, we can conceive of any "doing" involved in the production process as an activity,[2] and we quite naturally can separate the activities into direct and indirect activities—those directly involved in the creation of the commodity as against those not so directly involved.

Should activities be viewed as continuous or discrete? For the time being let us hold this question in abeyance and assume that all the activities are discrete in nature, that there are natural beginnings and ends to activities. Now suppose that the two processes P and R require only direct labor and the raw material, each activity is distinct, and the *number* of activities and their nature are exactly the same in the two processes. In this case a complete lack of specialization would be characterized by each person's carrying out every one of the activities involved in the process. On the other hand, the highest degree of specialization attainable would be that in which every direct worker carried out one and only one of the activities. We can also think of intermediate cases involving higher and lower degrees of specialization. For example, if in one process each person carried out, let us say, six types of activities whereas in the other each person carried out three types of activities, then we might call the latter process more specialized than the former.

There is a very important proviso to be added to the example in the last paragraph. In order to be able to say that one process is really less specialized than another, we have to be able to match individuals in the first process with individuals in the second (who carry out some of the activities of those in the first) and show that there is some individual in the first process whose activities embrace all of the activities of the associated individual in the second.

---

[2] It is to be noted that this is *not* the concept of an activity used in mathematical programming. Cf. Tjalling Koopmans, *Three Essays on the State of Economic Science,* McGraw-Hill Book Company, New York, 1957, pp. 72 ff.

This last point can perhaps be seen more clearly if we consider an example in which it would *not* be possible to say whether P is more or less specialized than R. Suppose that in process P each individual carries out three activities, designated as *a, b,* and *c* for the first group of individuals and *d, e,* and *f* for the second group. In other words, in this process two groups of workers divide the activities between them. Now suppose that under process R there are three groups of workers; one group carries out *a* and *b,* a second performs *c* and *d,* and a third does *e* and *f.* Can we say that R is the more specialized process? Obviously the first and third groups of workers under R are more specialized than their counterparts in the other process, and we can easily find their counterparts in P. However, what can we say about the second group of workers in process R? This group carries out a bundle of activities that is *not contained in* the bundle of activities carried out by any one group of workers in P. But suppose that the two really important activities in the production processes are *c* and *d,* and that the other activities are of an exceedingly trivial nature on the basis of any criterion we care to use. (We consider later what such criteria might be.)

Now following our supposition about the relative importance of the different activities, it would seem that the really important consideration in determining the relative degree of specialization would be the extent to which activities *c* and *d* are specialized. From this point of view process P appears the more specialized since in P the important activities *c* and *d* are performed by different individuals, and in this sense specialized, whereas this is not the case in R. Thus even in the highly simplified situation in which there is only one type of factor that is said to carry out the activities, and only one raw material, we can easily think of cases where it becomes necessary to weigh the relative importance of the activities involved if we are to draw any conclusions about the comparative degree of specialization of the two processes.

In all of the cases that we shall review there is at least this significant consideration: Is the bundle of activities carried out by each individual in one process *contained in* the bundle of activities carried out by some individual in the other process? This principle is equally applicable to the case of continuous activities, that is, to those situations in which there are no natural divisions between one activity and another, and where the separation of larger activities into smaller units is an arbitrary matter. As long as in one process the activities of each individual are contained in the continuum of activities of some individual participating in the other process, we can say that the former process is the more specialized, or at least that it is as specialized as the latter one.

THE "AVERAGE DEGREE OF SPECIALIZATION"

We consider now those situations where it is not possible to apply the "contained in" principle. Suppose that in process P some individuals carry out a greater number of activities than their counterparts in R but other individuals carry out a smaller number of activities than their counterparts in R. Strictly speaking this is an indeterminate case. But fruitful comparisons can be made in many situations of this sort that suggest, at least in a rough way, the greater or lesser degree of specialization involved. That is to say, we can think of means, in many cases, of setting relative weights to indicate the relative importance of different activities, and by summing the results we may arrive at the relative *average degree of specialization* that characterizes one process as against another. There are at least two problems involved here: What is the relative importance of different activities, and what is their relative "size"? For the present we assume that all activities are distinct and equal in size. Later in this chapter we shall examine the relative size question.

It may be possible to evaluate activities[3] in terms of the value that each activity adds to the product, or in terms of relative cost. Consider the value-added approach. Each activity transforms the raw material to some degree. This transformed piece of raw material may itself be looked upon as a commodity which has a market value, although it is an unfinished good from the consumers' point of view. In principle, the market would establish different prices for unfinished goods, and by comparing the activity that transforms one unfinished good into another with their prices we can impute a value to the activities so performed. Having done this, we can employ such relative valuations as our relative weights for the different degrees of importance of the different activities. By applying these weights to the two processes being compared, we determine whether on the average one process is more specialized than the other.

---

[3] We assume that the activities are the *same* in the two processes. We con-

The procedure just outlined may lead to an index number problem very similar to that found in income comparisons or in wealth comparisons: If our two processes are P and R, should we use the relative valuations determined on the basis of process P or those derived from process R? Will the two processes lead to the same valuation per activity or different valuations? A possibility along these lines would be the situation in which the *bundles* of activities are not the same in the two processes, although the *components* of all the bundles of activities are the same in the two processes. Suppose, for example, that processes P and R are composed of six activities, *a, b, c, d, e,* and *f,* but that in P the activities are divided into three bundles, *ab, cd,* and *ef,* while in R they are divided into two bundles, *abc* and *def.* We now have a variety of unfinished goods made of the raw material and the characteristics imposed upon it by the five bundles of activities. Consider for a moment how we might evaluate activity *c.* We can, in one instance, take the value of the good-in-process produced by the performance of activities *a, b,* and *c* as a bundle and subtract from it the value of the good produced by performing the bundle of activities *a* and *b.* This will give us an imputed value for activity *c.* On the other hand, we

---

sider later the complications that arise because activities are different in the two processes.

can take the value of the good created by performing activity bundles *cd* and *ef*, and deduct from it the value of the good created by performing the bundle of activities *d*, *e*, and *f*. This too will give us an imputed valuation for activity *c*. Should the valuations for *c* in the two cases turn out to be exactly the same, then no special problem arises. But it may be that in some fashion or other activities *c* and *d* are highly complementary and for that reason having a commodity that contains *c* in combination with *d* is worth more than having one that contains *c* with other activities. In this case it is clear that we have a problem in determining what value of activity *c* we should use in assessing the average degree of specialization of the two processes. If, when we use either of the two valuations for the activities, it turns out that in both cases one process is on the average more specialized than the other, then, of course, we ought to refer to that process as the more specialized one. If, on the other hand, we discover that with one set of relative valuations of the activities we get one result and with another set of relative valuations we get another result, then we have an indeterminate case.

The problems just considered become somewhat more complicated once we add indirect labor or indirect equipment. Consider indirect labor. The immediate difficulty is that there is no way of determining how much the indirect labor adds to the value of the product, since the indirect labor itself does not directly contribute to changing any of the characteristics of the raw material, and as a consequence we cannot impute from the increased value of the raw material the value of indirect labor activities performed. Of course, there is the possibility that each bundle of indirect activities in process P is contained in a bundle of indirect activities in process R. This is the case of greater absolute degree of specialization for process P with respect to the indirect activities. If it turns out that P also has a greater average degree of specialization than R, by whatever scheme of evaluation of the direct activities is employed, then clearly there is no special problem, and we would consider P the more specialized process on the average. However, suppose it is R that has the greater average degree of specialization with respect to the direct activities. In that case if we are to persist in trying to determine which process is on the average the more specialized, we must reëvaluate all activities by some other means.

As previously suggested, we can attempt to assess the relative importance of activities in terms of their relative costs of performance in each process. (This is only a suggestion, and we will not dwell on all the complications that may arise in trying to carry it out in detail.) The reader will note that it is quite possible under this procedure for activities in one process to be evaluated differently from the same activities in the other process. We have seen that under some circumstances this difficulty can lead to indeterminate results, but it certainly need not in all cases lead to indeterminate results. In other words, there may be a large class of cases in which the result is the same for either set of relative evaluations of the activities.

THE ACTIVITY UNIT PROBLEM

We may think of the degree of specialization as the inverse of the average number of activities[4] performed per man, or per acting agent. Roughly speaking, the greater the number of activities performed per man, the lower the degree of specialization. However, behind such a conception lies the problem of finding at least roughly equivalent activity units. Often there are both big activities and little activities in the same process, and all of the activities within a bundle (a set of activities performed by a single acting agent) should not be penalized because of the addition to that bundle of, say, a very little activity. For example, suppose that process P contains three activity bundles of two activities each, and suppose further that the activities in the first two bundles are quite substantial whereas the activities in the third bundle are exceedingly small and require exceedingly little time to perform. Now suppose that we create a new process—call it process N—by dividing up the activities in the third bundle of P and adding them to the activities in the first two bundles. In view of the limited nature of the activities contained in the third bundle of P, the new process N would appear to be hardly less specialized in any important sense than the old process P. However, if for the new process N we assign a degree of specialization that is the inverse of the average number of activities per acting agent, then N will have a much lower index of specialization than P. This appears to be an important difficulty since it goes counter to the reasonable principle that we should not permit inconsequential aspects of a situation to change markedly our general evaluation of the situation in its totality. Suppose that there exists a third process, R, whose index of specialization is only slightly above that of P. Now, process N, with the two activity bundles, may turn out to have an index of specialization considerably below process R's, but since it is hardly less specialized than the first process (P) this would seem a contradictory result.

---

[4] Assume for a moment that these activities are each of equal value.

The remedy must lie in finding comparable activity units. What precisely would be the appropriate measure of the relative *size* of an activity is difficult to say in the abstract, and in individual cases the measure used must be, to some extent, arbitrary.[5] With respect to activities contained within a bundle of activities, or a smaller activity contained within the continuum of a larger activity, some concept of average time required for performance would probably be the best measure. If it turns out, in any case, that there is some more convenient or more suitable dimension, there is no reason why we should not use that in specific instances in order to assess how small we deem a small activity to be in comparison with a large one.

SUMMARY AND CONCLUSION

The foregoing discussion suggests that it is not always meaningful to say that one production process is absolutely or on the average more specialized than another. The difficulties of comparing the relative average degrees of specialization of different methods may be due to any one or a combination of the following: (1) the difficulty of determining equivalent activity units; (2) the difficulty of assessing the relative importance of different activities within a process; and (3) the fact that some activities in one process may not have any counterpart in another process.

---

[5] We distinguish between the size of an activity and its value. A very small activity may add a great deal to the value of the output, or vice versa. Clearly, the relative size of an activity and its relative importance are two different matters.

Some production methods may be only partially comparable.

1. For example, processes P and R may contain some activities which are the same and some which are different. For the activities that are the same, the activity bundles[6] in P may be contained in at least some and possibly all of the activity bundles in R. In this case, for the portion of the processes that are comparable, we can say that P is more specialized than R in an absolute sense.

2. Another possibility is that some of the activity bundles of P may be contained in some of the activity bundles of R, and also that some of the activity bundles of R may be contained in some of the activity bundles of P. In this case we must judge the relative importance of the two segments in terms either of value created or of cost if we wish to say anything about the difference in average degree of specialization.

3. For all comparable activities the activity bundles of P may be contained in the activity bundles of R, but the incomparable activities may be known to arise entirely because of the degree of specialization of process P. In other words, in the case where we can clearly attribute the activities that are different in the two processes entirely to the differences in specialization, we can say that one process is more specialized than the other.

4. There may also be cases in which the indirect activities cannot be attributed to different degrees of specialization, but in which the indirect activities are nevertheless comparable in the two processes. If for process P both the direct and the indirect activities are more specialized than those in R, there is of course no problem in determining which is, on the whole, the more specialized process, but if a greater degree of specialization exists only for one category

---

[6] Once again by this term we mean the set of activities performed by a single acting agent.

of activities and not for the other, then some relative evaluation must be made between the direct and indirect activities.

What is the upshot of our discussion?

The concept of relative degree of specialization is a useful one. But it probably has to be treated with the same care as such concepts as income or wealth. It seems clear that the concept can be used because a set of cases exists, perhaps a large set, in which there is meaning to the notion that one process is more specialized than another. The important point here is that in many situations an assessment about differences in the degree of specialization must involve judgments as to the relative importance and relative sizes of different activities. But a clear-cut comparison is not possible for all pairs of processes. In addition, when we do use the notion of greater degree of specialization we should keep in mind whether it is used in the absolute sense or in the average sense in terms of cost or value created in order to have a definite idea as to the meaning involved. We have also seen that because of the index number problem involved there may be cases where the processes are in principle comparable, but where in fact we cannot say whether one process is more specialized than another. In such a case we cannot deduce that the two processes are equally specialized. It is simply that there is no way of knowing which is the more specialized of the two. Finally, we must remember that, for the purpose of understanding the reasons for differences in the efficiency of different processes, specialization is not the only factor. One process may be more efficient than another for reasons other than the relative degree of specialization involved.

8

Some Determinants of the Degree of Specialization

THE ADVANTAGES OF SPECIALIZATION

Underlying our discussion of costs and cost curves in previous chapters are the notions of economies and dis-economies of scale. These are usually attributed in part to the different opportunities for specialization that arise with different scales of operations. Labor and other factors perform their productive activities more efficiently when used in a specialized manner. In addition, the specialization of tasks, in order to be effective, requires coördination, which in turn is one of the central elements of the more general problem of organization. Hence we consider here the elements that determine the limits to which specialization can be carried out effectively. Much of what there is to be said on this topic has been treated by a number of writers, and some of the arguments go back, of course, to Adam Smith's first chapter in the *Wealth of Nations*.[1] But some of our discussion of these matters will be from a slightly different viewpoint from that usually found in economics or management books.

---

[1] The classical economists frequently discussed this question, and their

Following the ideas developed earlier we would say that if two processes are different only in their degree of specialization, and if they are truly comparable, then the main differences between them are the activity bundles and perhaps the resources related to the nature of the activity bundling, but not the activities themselves. To get at our problem let us examine the advantages and disadvantages of different ways of clustering activities. By going to the point where any additional advantage that arises from an increase in the degree of specialization is just offset by the consequent disadvantages thereof, we get at the limits to the degree of specialization from an efficiency standpoint.

Many activity bundles can be broken up into smaller bundles, and when the smaller bundles are assigned to different acting agents a greater degree of specialization is achieved. But, of course, there is little reason to have specialization simply for its own sake. If we are to have a higher degree of specialization it must be because there are advantages therein.

1. *The allocation advantage.* The greater the specialization, the easier it is to allocate the factors of production to those activities to which they are best suited. This very old idea is one of the cornerstones of international trade theory, where it appears in connection with the concept of comparative advantage. The logic of the allocative advantage can perhaps best be seen from the point of view of an agent who is to carry out the activities under analysis. The main assumption is that any agent cannot perform all possible

---

treatises often contained a chapter on the division of labor and specialization. See, e.g., John Stuart Mill, *Principles of Political Economy*, Longmans, Green, New York, 1920, pp. 122–130, or J. B. Say, *Treatise on Political Economy*, Wells and Lilly, Boston, 1824, chap. VIII. Charles Babbage, who according to Marshall anticipated the ideas of scientific management, has a very extensive treatment of the problem in his *Economy of Machinery and Manufactures*, John Murray, London, 1846, chaps. XIX and XX. The neoclassicists seem to have added very little to the discussion, and we find only scattered references in both Marshall's *Principles* and *Industry and Trade*. Modern economists seem to have lost interest in the question, although older texts, such as F. W. Taussig's *Principles of Economics*, 1939 edition, The Macmillan Company, New York, 1939, do have a chapter on labor specialization.

activities equally well. Suppose that we list, in order of proficiency, all of the activities that an individual can perform. If he is to perform all of the activities on the list, his efficiency will obviously be smaller than if he performs only the activity that is first on the list. With a lesser degree of specialization an individual would have to perform a greater number of activities, and hence he would have to perform some activities that are lower on the list than would be the case with a greater degree of specialization. The greater the degree of specialization, the narrower the range of activities performed by any one agent of production, and the *easier* it is to allocate activities in such a way that each person is performing only those activities nearest to the top of his list.[2]

The allocation advantage may be just as applicable to machines as it is to individuals. Compare the case of a multipurpose machine with that of a single-purpose machine. Since the multipurpose machine cannot do everything equally well it would pay to use it in a specialized manner for the specific use to which it is best suited, and it would pay even more to design a machine for only that best use without wasting the additional resources necessary in order to incorporate the additional uses into the machine. Indeed this point deserves to be emphasized. Probably a great deal of the historical increase in productivity per man that has taken place is in part attributable to the fact that specialization of labor makes possible the design and creation of exceedingly productive equipment and machines. In other words, rather than point up the division of labor as an effective scheme by which to achieve high productivity we should emphasize that the possibility of the division of labor is what enables us to develop the highly specialized machines that make some modern output procedures so effective.[3]

2. *The direct efficiency advantage.* Entirely apart from the allocation advantage, an activity may possibly be performed more efficiently when it is specialized than when it is carried out in

---

[2] It may not be possible to carry this ideal into practice in every respect. Some activities may be low on everyone's list, but they may need performing by someone nevertheless.

[3] Cf. A. Marshall, *Principles of Economics,* 8th ed., Macmillan and Company, London, 1924, pp. 255–257.

conjunction with other activities. That is, the same person or agent carrying out a certain activity by himself and repeating it endlessly may do it more efficiently than if he interspersed this activity with a number of others. For example, it is not uncommon for a person engaged in a given task to have a warming-up period during which the quality and speed of his performance increases. The same thing may be true for certain types of machinery which perform more efficiently after they are warmed up. The example just mentioned must be distinguished from another possibility, namely, the possible advantage of sustained practice. That is, with a specialized activity the person carrying out the activity gets so much sustained practice that his performance is higher than it would be without that practice.

Some activities in a production sequence operate in the manner of overhead costs. With respect to many types of activities we may have to set the stage for performance or initiate performance, as well as do various things to terminate performance in a desirable manner. Now a possible advantage of specialization is that we may avoid the costs of starting and stopping when we do not have to terminate one activity in order to begin another. By lumping many units of the same activity together we spread what may be looked upon as the overhead starting-stopping costs over a great many units, rather than having each unit bear the entire burden.

3. *The investment, training, and adaptation advantage.* In many production processes a person may have to be trained or adapted to his new tasks and working conditions. It seems reasonable to expect that if the nature of the job involves a series of varied activities it will take longer to train and adapt a person than if only one or relatively few activities are involved.

The cost of investing in increasing performance of an activity and the returns obtained thereby may differ whether the investment is made in one activity by itself or in a bundle or sequence of varied activities. This is likely to apply to men as well as to machines. For example, consider the alternatives of educating a man to a certain degree to carry out a specialized activity as against educating him to carry out a series of activities of which the former specialized activity forms a part. There are two elements involved here: to what degree should he be educated, and what are the relative returns obtained from educating him with respect to the specialized activity as against the bundle of activities? In some cases it may not pay to give a high degree of training to an individual to carry out a series of activities, whereas it may well pay to train him to carry out a single activity. Human capacities are limited, and therefore the training necessary to carry out the bundle of activities may be more than the sum of the trainings necessary to carry out these activities by as many individuals as there are activities.

SOME LIMITS TO SPECIALIZATION

One of the main reasons why the degree of specialization is not extended indefinitely is that in most cases a very high degree of specialization can be carried out economically only if the scale of operations is rather large, or at least beyond some minimum.

In most production processes there is a fixed size, or at least a minimum size, to many of the components that enter into the production process. A component may be naturally indivisible, or it may come in a fixed size or in relatively few fixed sizes. If a factor of production is indivisible, then it usually pays to use that factor to its full capacity and certainly to its optimal capacity, where full capacity and optimal capacity do not coincide. For example, if a particular factor of production can be used in the carrying out of 100 units of an activity per day, it will not pay to employ that factor to carry out only a single unit or very few units of the activity per day. In small-scale operations it may pay to crowd many types of activities to be performed by a single factor of production; otherwise it will be idle part of the time. If the factor did not have a fixed size, this would not be an issue since we could then simply employ a portion of the factor in as specialized a manner as desired and use other portions in other specialized ways.

The fixity or indivisibility of an element in the production process may lie either in the commodity, in some of the specifications of the commodity, or in the specifications in some of the components of the commodity. For example, if the commodity is for human use, then clearly it must fit certain specifications connected with the size of human beings. It would not do to make chairs only half the size appropriate for the human anatomy. In this case, for example, the specialized machinery conceivable for chairmaking may come in any size whatsoever, and therefore be said to be divisible, but the machines that can be utilized efficiently will be limited to those consistent with the specifications of man-sized chairs.

The fixedness or specificity of size of a component may also lie in the raw material or in some part of the raw material, especially where the raw materials involved are of a biological nature. Most animals used in food-processing or in animal products come in only a narrow range of sizes. As a consequence the activities which form the production processes involved must also be of a size appropriate to the raw material being processed, and therefore the activity and the factors of production used in that activity will also have to be of some minimum size. Thus we can see that there must be a minimum scale of operations if the activity is to be carried on efficiently in a highly specialized manner. Otherwise, with highly specialized equipment or personnel and a very small scale of operations, a specialized factor will be forced to be unemployed a good portion of the time.

To some extent size may be determined by the specification of the user although the commodity and the factors involved may all be infinitely divisible. For example, milk does not come in 10-ton containers, or in containers of one gram, although both might be feasible on purely technical grounds. Obviously, the limited capacity of the human agent, as well as such factors as the cost of storage and the trouble of combining a very large number of small units, has something to do with it.

In this connection it may be worth while to point out that the geometry of the case may in part dictate the economical size of an element in a production process, although any size may be technically possible. For example, it takes proportionately less glass to build a container to hold a thousand cubic centimeters of a liquid than to build one to hold only one cubic centimeter. As we increase the volume of the liquid to be held in the container, the glass required for the container will not increase proportionately. There may be other reasons, of course, why we would not want containers of too large a size, suggested by such things as storage costs, perishability of the commodity, etc. Taking both sets of considerations into account, we may easily see that some optimum size container exists. And such an optimum, usually set at the nearest conventional measure, will determine the fixed size of the container, or of some other component in the production process, although any size might be technically feasible. Other examples of this type are usually all around us in everyday life and hence need not be enumerated. Of course, in most cases the fixity of size of a component of the production process will be determined in part by man-made decisions and in part by natural fixities and indivisibilities of some of the factors involved. The point is that even if such natural indivisibilities or fixities did not exist, there might still be reason to establish a fixed size or a limited set of such sizes for many production and consumption components.

Two more factors must be mentioned as possible limits to specialization: monotony and inventory costs. If specialization is carried to a very high degree there may be a loss of efficiency through monotony of the manual operations.[4]

In a world of uncertainty inventories are necessary. They are often necessary for every stage of the production process that involves parts, or raw materials. The greater the number of stages into which the production process is subdivided, the greater the inventories needed. In the face of indivisibilities, inventory costs per unit are not proportional to output. Up to a point the larger the run, the lower the inventory costs. Hence, at an insufficiently large scale, excessive specialization can involve very high inventory costs per unit. The inventory cost aspect should be kept in mind at all stages of our discussion.

---

[4] On this question see Elton Mayo, *Human Problems of an Industrial Civilization,* The Macmillan Company, New York, 1933, Chap. 2.

STANDARDIZATION VERSUS VARIETY

One of the usual accompaniments of a high degree of specialization is standardization. A high degree of specialization implies the repetition of a minor activity in a sequence over a long period of time. There are two choices here: the repeated activity may be precisely the same time after time, or it may be permitted to differ so as to take into account differences in detail in the final product. Obviously the second alternative is likely to mean the abandonment of many advantages of a high degree of specialization. If the activity is to differ time after time, then the worker and his machines have to adjust themselves or be adjusted to the differences in the activity. But it is precisely such adjustments which are required in the shift from one activity to another. In other words, the advantage of specialization under standardization is that the details of adjusting machinery, or making decisions as to how to perform a certain act, can be avoided when the activity is one that is repeated endlessly. From this it can be seen that if the advantage of specializing an activity is to be retained, a high degree of standardization is usually called for. Furthermore, the product has to be standardized, not as an end in itself, but as a consequence of utilizing specialized activities and gaining the advantages of specialization.

Another reason for standardization is complementarity between components. If a commodity has three parts, say A, B, and C, the specifications of A cannot usually be determined separately from those of B and C. Examples are numerous, that of nuts and bolts being a rather obvious one. As a consequence it is desirable that components of different units of a commodity be interchangeable. Standardization of components, and hence standardization of the commodity itself, is the means of achieving such interchangeability.[5]

To see the other side of the coin we might touch briefly on the costs of introducing variations in a commodity and its components.

---

[5] For specific examples to illustrate the importance of standardization in the history of technology, in order to achieve low-cost, high-quantity production, see Abbott Payton Usher, *A History of Mechanical Inventions,* Harvard University Press, Cambridge, 1954, pp. 375 ff.

These include the costs of starting and stopping, and the costs of readjusting activities and machinery in order to achieve slight changes in specifications. What is lost is the advantage to be obtained from the persistent repetition of a single activity or narrow group of activities. We have already indicated some aspects of this advantage. For example, the ease of learning a certain operation will clearly be facilitated if the operation is completely standardized. Otherwise one has to learn simultaneously a number of variations on the same operation which may require a higher level of native talents as well as a greater amount of time. Similarly, investment in machinery that performs in a completely standardized manner must be less than investment in machinery that permits a number of variations.

But standardization is the enemy of variety. This fact has both positive and negative aspects. For example, standardization simplifies choice. Although, other things equal, the greater the variety of goods obtainable, the greater the utility obtained from them, still variety introduces complications into the process of choosing. There are at least two costs to be considered. First, there is the information cost—the cost of discovering the qualities and specifications of each of a large number of variations on a single product. This is especially important with respect to durable goods, where we have to be able to tell not only how the commodity will work immediately but how it will function over a long period of time. Second, there is the cost of having to consider a large range of alternatives when making a choice. The problems of choosing are likely to increase with variety to a greater degree than the increase in the variety itself. If a commodity has several properties and we have to choose only between two alternatives, then only one comparison between the various properties has to be made. But if the choice is between three alternatives, A, B, and C, then three pair-wise comparisons must be made, between A and B, A and C, and B and C. In a similar vein it can readily be seen that the number of pair-wise comparisons will increase proportionately more than the additions to variety, and it follows that in some cases the additional utility gained from a wider choice is less than the cost of making the choice out of the greater variety.

But there are, of course, limits to the degree of standardization and the narrowing of the range of choice that producers can impose and still retain adequate demand for their product. Anything that limits the degree to which products or components of products can be standardized also limits the degree to which productive activities can be standardized, and hence limits the degree of specialization. The producer must balance the greater value of his range of products if he offers some variety against the greater cost of producing them because of the reduced degree of specialization that follows therefrom.

COMPLEMENTARITIES BETWEEN VARIOUS ACTIVITIES

In the previous sections we have stressed the limitations to the extension of the degree of specialization that arise from an insufficiently large scale of operations. However, there may be certain conditions surrounding the nature of the factors, or the nature of the commodities, that may limit or make unwise an extension of the degree of specialization entirely apart from scale considerations. One such situation may arise where the quality of performance is closely connected with the coördination of different activities or the relation between different components of the commodity. For example, in some delicate pieces of machinery the different components have to be connected to each other so that they fit within very narrow tolerances. In that case the activities may have to be carried out in a sequence by the same individual, or by the same piece of machinery, because it may be easier to establish the exact fit between components if the same individual or machine produces both components than if they are done separately. In addition, should there be some slight errors made in producing one of the components, it may be easier for the same individual to make the compensating adjustment in the other component than to have one individual produce one component and a second individual check the errors and then work out the compensating adjustments.

Another situation limiting the advisability of extending the degree of specialization beyond some point occurs when someone or something performing activity A can perform activity B simultaneously at no extra cost or at little extra cost. For example, a machine that, let us say, dries a wet surface may at the same time, and with only a small additional component, also polish this surface immediately after the drying operation. A more homely example: suppose a man does something that requires only the use of his right hand. If a related activity can be found that utilizes only his left hand without interfering with the right-handed activity, it may be economically desirable for him to perform both activities simultaneously. In general, we have to balance the advantage of simultaneous motions or activities against the advantage of specialization.

It is of interest to note that physical indivisibilities of a factor of production, or of a component of the commodity, may work in two ways to limit the degree of specialization: Indivisibilities may require that a certain minimum scale be achieved before the use of the indivisible factor will be economical. Apart from the scale aspect, the existence of an indivisibility may permit the performance of some activities simultaneously at little or no extra cost when performing only one of these activities.

THE DEGREE OF SPECIALIZATION UNDER SEQUENTIAL PROCESSES

The point to which specialization is carried may be limited by the nature of the production process. Most, perhaps all, processes are at least in part sequential processes, i.e., at least some of the activities involved have to be carried out in a given sequence.

We can distinguish between sequential processes or segments thereof that are stoppable and those that are not. By a stoppable process we have in mind one in which it is possible to stop the sequence of activities and do something else before completing the sequence. For example, surgical operations, or dental operations such as pulling teeth, are usually nonstoppable processes. Personal services frequently fall into this category. In some cases it may not be physically feasible or technically desirable to stop the process. In other cases it may be both physically feasible and technically possible but not economically desirable to do so.

What is the significance of nonstoppable sequential processes with respect to specialization? If the entire process is nonstoppable and sequential and if it can only be carried out by a single individual or single acting agent, then there is no opportunity to specialize. If the entire sequence has to be repeated for each unit by an acting agent, and if it cannot be stopped before completion, then the only thing to do is to produce each unit separately. There is no possibility of obtaining the advantages that arise from the repetition of minute activities. This may perhaps explain why some types of personal services lag behind with respect to productivity while other productive activities become more efficient in the course of time.

Where the process is sequential and nonstoppable, but where it need not be carried out by a single individual, it may be possible to use specialized agents for various portions of the process and in this way obtain some of the allocation advantage of specialization. The degree of specialization would be limited by the extent to which the sequence could be separated into stages and carried out on a "relay team" basis. Clearly, many of the other advantages of specialization could not be obtained in sequential nonstoppable processes.

Where the process is completely sequential but stoppable a greater degree of specialization is possible and more of the advantages of specialization are obtainable. For example, if a sequential process is stoppable, then given a sufficient scale of operations one could take advantage of the uninterrupted repetition of small segments of the process.[6]

Consider now variable as against proportional sequential processes. By a proportional process we have in mind one in which it is necessary to carry out the activities in the sequence in proportion to output. Thus, if ten units are produced, each activity is carried out ten times as much as it would be if only one unit were produced. In a variable sequential process the activities still have to be carried out in a given sequence but not in proportion to output. Strictly speaking, the latter is not really a single process as we have used the term—it is a set of processes in which all members of the set have the same activities in common and the same sequence in common, but the frequency with which the same activity is carried out may differ for different members of the set.

---

[6] These remarks are in accord with the emphasis in later chapters on the organizational and economic importance of repetition and duplication, especially in the face of incomplete knowledge. Inventory costs may be of importance in this connection.

The greatest possibility for cost reductions with respect to increases in scale is likely to be offered by variable sequential processes. There are several reasons. In the first instance it is possible to obtain all of the advantages of specialization with respect to scale that were possible in the case of the proportional (and stoppable) sequential production process. Cost reductions also accrue as output increases because certain activities, or activity segments, do not have to be repeated in proportion to output.[7] Now, the ultimate reasons why a process should have this characteristic are not easy to determine. But we can suggest some possibilities. What is involved is something that might loosely be called the "stamp effect." That is, the outcome of later activities is determined by the outcomes of the earlier activities. For example, a mold or die used in a production process will "stamp" on the results of the later activities the form determined by the earlier ones involved in making the mold or die. Another way of looking at it is to suggest that between the elements of such processes there exists a sequential complementarity, i.e., the characteristics of the results of later activities in the sequence are fixed, at least in part, by the nature of the earlier ones, and in addition the outputs of earlier activities have at least some degree of durability. Since the results of these earlier activities are durable, and since these processes are such that they utilize in the later activities the outputs of earlier ones, it follows that the later activities can be carried out without the necessity of always repeating the earlier ones. The outcome of a set of earlier activities may be tangible or intangible. It may be a die, a mold, a piece of equipment, and so on. Or the result of the earlier activities to be used later may be a design, a plan, an invention, or an idea. In the latter case it is intangible and infinitely durable.

---

[7] In this connection it is perhaps important to note that there may be no optimum proportions of factors. That is to say, the greater the output, the lower the cost of the nonvariable or less variable activities per unit of output. This point is of some significance in connection with the "proportionality controversy." One version of this doctrine is based on the notion that for some output there exists an optimum combination of factors, and that the same combination could always be achieved for greater outputs. But it may very well be that no such optimum exists. See E. H. Chamberlin, *The Theory of Monopolistic Competition*, 6th ed., Harvard University Press, Cambridge, 1948, Appendix B. Also, H. Leibenstein, "The Proportionality Controversy and The Theory of Production," *Quarterly Journal of Economics*, November, 1955.

Where all of the production processes are not of the variable sequential type it may be possible and desirable to shift from a proportional sequential process to a variable one as we increase the scale of operations. In other words, for a low scale of operations it may pay to employ the rigidly proportional process whereas for a higher scale it may be economical to use a variable process. This is likely to be the case where the invariable segments of the variable process are relatively expensive compared to their counterparts in the proportional process. Consider the processes of designing and then manufacturing an article. The design component is likely to be expensive if the production process is to be carried out in a highly specialized fashion, since it is necessary to design the components of a commodity in such a way that they readily fit one with the other. (On the other hand, if all activities have to be repeated in the production of each unit so that the fit can be corrected as we proceed in manufacturing one component after another, then the necessary design element in the production process need not be as carefully carried out, and in some cases it may never be carried out in a formal manner.) If the variable sequential process is to be used, the scale of operations must be sufficiently large to pay for the more expensive design component.

Other examples of variable sequential processes involve the introduction of inventions, the production of automobiles on an assembly line, and in general most processes that involve capital creation or an increase in the capital components of production.

For example, consider innovations. Clearly there is considerable effort, as well as cost, involved in shifting from one production process to another. However, once the shift is made and the effort expended we can continue with the new process indefinitely. Thus, the cost of introducing an innovation is in the nature of a fixed cost, and the new process must be looked upon as a variable sequential process. The scale of operations, or the period of time during which the new process will be employed, must be sufficiently large or long to pay for the cost of introducing the innovation. Similarly, any act of capital creation involves shifting to a production process in which first the additional capital equipment is created and then the production of the commodity itself proceeds.

As a transition to the chapters that follow we may keep in mind that the introduction of an organizational innovation can be viewed as a shift from one variable sequential process to another. Furthermore, the setting up of an organization, or the reorganization of a firm, is an element of a variable sequential process. In some cases a limit to the extension of the degree of specialization on economic grounds may arise because, for the anticipated scale of operations, the costs of the greater degree of organization required by the increase in specialization may outweigh the advantages of specialization.

Part III

Organizational Analysis

9

Roles and Role-Taking in Organizations

THE ROLE APPROACH TO ORGANIZATION THEORY

In any attempt to develop the idea of an organization it is necessary to begin with some basic components. Sometimes the problem is broken down in terms of the basic activities that the organization has to perform, or in terms of the specific officers or categories of officers that may be said to control the organization.[1] Our analysis will, at the outset, be carried out on a slightly higher level of abstraction, but not the highest level possible. Somehow in developing an organization theory one has to choose an intermediate level between considering things on so general a plane that it seems to have very little content and considering them on so specific a level that it boils down to a small category of cases having exceedingly limited scope. The level aimed at here is an intermediate one between these extremes.

Our basic unit of analysis will be the role. We shall look upon and attempt to analyze organizations in terms of the relations between the roles that make up the organization. What is being asserted here essentially is that a human organization is an amalgamation of certain roles necessarily in some relation with each other. For our purposes this is a useful way of looking at the problem, although it may not be the only useful way of doing so.

---

[1] This is essentially the organization-chart approach.

When we come to consider the firm we shall look upon it as a subset of that class of entities that we call organizations. Principles or generalizations that we shall establish for organizations as such will also be applicable to firms. Further, we shall look upon roles, to be examined carefully and defined in a special sense in this chapter, as the key abstract units of an organization. Specifically, when we think of the organization at work we shall have in mind specific role-takers making certain decisions and performing certain tasks related to these decisions.

The suggestion that we employ the concept of role in organization theory has to be defended. Professor Simon, in the second edition of his *Administrative Behavior*,[2] criticizes the use of what he views as the sociologists' concept of role, which he takes to mean a pattern of behavior. I agree entirely that the concept of role defined simply as a pattern of behavior is inadequate for the purpose of developing a theory of organization. As the following pages will make clear, the notion of a role to be used here is very much broader.

For Simon, decision-making is central to the problem of organization. It is also central to the concept of an organization that we shall develop. But decisions are not made by individuals with entirely undefined parts to play in an organization. In other words, it seems quite natural to view decisions as being made within the roles played by individuals who are members of an organization. Role is, after all, a fairly common word. In the social sciences we are often forced to use common words in a special sense. From that point of view it seems quite useful to attempt to define role in such a way as to retain the concept but avoid the implications of Professor Simon's strictures.

---

[2] Herbert A. Simon, *Administrative Behavior*, The Macmillan Company, New York, 1957, pp. xxx ff.

The following analysis will be tied in many ways with the ideas elaborated upon in the previous chapters. For example, we shall visualize the role-taker performing various activities, and these activities will be specialized and come in bundles. Thus we have a connection with our previous discussions of activities, specialization, and efficiency. Indeed, it is probably impossible to analyze human organizations without to some extent considering the specialization of activities and some of the efficiency considerations involved therein.

Tentatively (and roughly) we may look upon an organization simply as a somewhat durable group of individuals that perform certain activities, at least some of which are in pursuit of common aims.[3] The question immediately arises: How are the common aims of the organization determined and what mechanism generates their pursuit? We would not expect a group of people simply doing things, so to speak, to necessarily find the results of their activities in any way fulfilling common aims. To pursue common aims some activities have to be coördinated. This implies that the interrelations between individuals must be of such a kind that coördination does take place. We are not concerned with the relationships between specific individuals. This would be looking at the question from much too concrete a viewpoint. Rather, given certain aims to be pursued or objectives to be reached or approximated, we are interested in knowing what relations would have to exist or be imposed between the members of any group of more or less *typical* individuals that would lead to the pursuit of some common aims. In other words, we are more interested in relations between roles than between specific individuals.

It is immediately evident that at least some activities which go on within an organization are not in pursuit of the aims of the organization directly but are necessary in order to coördinate the various direct activities.

Consider briefly some of these organizational activities.

1. Somehow a decision has to be made on the structure, scheme,

---

[3] This should not be taken as our final definition of an organization.

and scheduling of the direct and indirect activities. The structure need not be formalized and precise, but in some way it must be decided, at least in part, *who is to do what.* The decision is not necessarily made by any one individual.

2. The specific individuals who are to carry out the various activities (i.e., the roles) have to be chosen. Whereas the first decision above has to do with determining the nature of the roles, which are essentially unfilled, now specific individuals must be assigned to these roles.

3. When operations begin, at least some of the role-takers and their activities must be directed, coördinated, and controlled. This, of course, really involves three different types of organizational activities, but we may list them under one heading for now because somehow they must be interwoven and carried out simultaneously.

4. At every stage organizational activities have to be carried on in the interests of maintaining the organization or determining the termination or the transition of the organization into some other form. On a day-to-day basis these activities may be unobtrusive, but they are most necessary if we are to think in terms of an organization rather than simply a gang of individuals carrying out a one-time-only task.

An obvious characteristic of the organizational activities we have just mentioned is that they involve decision-making as well as performance activities. Indeed, the decision-making activities are likely to strike one as the more important of the two types of activities. For example, the choice of actors to fill the roles is clearly a decision-making activity. The determination of the specific nature and structure of the roles to be played is also of a decision-making nature. Similarly, we can see as we proceed that the other types of organizational activities have significant decision-making components. Clearly, if we are to understand the functioning of an organization, we must simultaneously understand the nature of the decision-making mechanisms within the organization. Accordingly we shall develop our concept of role so that roles and the relations between roles embrace the decision-making mechanism.

THE NATURE OF AND *RAISON D'ÊTRE* FOR ROLE-TAKING

The reasons for role determination and role-taking in human society lie, in part, in two characteristics of the nature of human activities. Briefly put, human activities are neither completely predetermined nor completely random. In other words, the mere fact of being an individual does not force one to carry out a specific predetermined sequence of activities, nor does it force one to do various things in an inexplicable manner. It pays to have part of our activities organized and purposive. The implication is that we do have some choice with respect to the nature of the organization of our activities, and that some organizations are, from various points of view, superior to others. In addition, role-taking is more efficient than complete non-role-taking.

To see this a little more clearly let us engage in the following *Gedankenexperimente:* Suppose we begin with a state of nature in which there were neither specific roles nor organizations. What would occur?

In this so-called state of nature that we start with, a variety of activities could be carried out. Some of these would be almost necessary; others, whether or not they were performed, would certainly be a matter of choice. Choice would exist not only with respect to whether certain activities were performed or not but also with respect to the sequence of activities and the bundling of activities into convenient clusters. It is out of this element of choice that the grouping of activities, including decision-making activities, and the playing of roles become possible.

Why some activities are bundled together and carried out as a group or in a sequence, perhaps for a certain part of the day or week or month, and other activities are left for other occasions is a complex question that depends in part on the objectives of the individual and in part on his training within the family and other social groups. If we are to see how some roles arise, we must assume that some objectives, however vague, exist. Assume that the objective of individuals is to fulfill their desires somewhat efficiently, given their resources and knowledge. Granted that this is not the only objective, and that even where it is an objective its effects can usually not be seen clearly, because it may be mixed simultaneously with other objectives, considerations, and constraints. However, let us see where we come out. We can attempt to build in our complications at a later stage. First, consider the situation of an isolated individual, say a Robinson Crusoe.

Our isolated hero is surveying the host of activities he can conceivably engage in and considering the problem of choosing among the alternative ways of organizing their performance. Of course, the mere fact that we assume he is surveying the situation and considering a decision in this matter already implies a minimal degree of organization of activities—a separation of the task of organizing routine activities from their performance. Some of the activities will clearly be of an everyday sort; others will strike Crusoe as of a special kind. It is likely that those activities which are to meet routine wants, or wants that are repeated day after day, will be decided upon and organized first.

Crusoe's problems of organization involve at least these two aspects of choice: what activities to perform and when to perform them. If certain wants are given and there are competing activities that can lead to the satisfaction of these wants, then a choice must be made between the competing activities or competing bundles of activities. This choice can be made on efficiency grounds—that is, the bundle of activities that will satisfy the wants most fully will be the one chosen—or on other grounds. When are the activities to be performed? This will depend partly on when the wants will appear most urgent, but to some extent the fulfillment of wants can be postponed and so a sequence of complementary activities may be carried out. For example, Robinson Crusoe will not do things just as they occur to him but will segregate different portions of the day for different types of activities. First, he may play the role of a food gatherer; then he may take on the role of a food storer; after that he may become a food preparer and following that a consumer. Other portions of the day may be allocated to the maintenance of his few possessions, to the creation of other possessions, and perhaps to keeping his diary. Clearly, this seems to make more sense intuitively than the alternative of doing various activities that fall within these groups on a random or helter-skelter basis. We can hardly conceive of his writing three words in his diary, running to pick up a nut, putting a stitch in an article of clothing, and then rubbing some twigs together in an attempt to start a fire but almost immediately dropping this for something else. Some activities can be carried out simultaneously to a greater advantage since they have complementary features, while for others their complementary nature may be taken advantage of by performance in an unbroken sequence. Broadly speaking, we may look upon these sequences of activities or activity bundles, in those cases in which more than one activity is carried out simultaneously, as examples of the playing of a role. But it must be emphasized that this is only part of the picture.

We may, if we wish, look upon this business of role creation as an act of exchange even for Robinson Crusoe. Our isolated individual can be said to be giving up the choice of doing whatever he pleases, whenever he pleases, for a more ordered and scheduled sequence of activities.

The determination of the sequence of activities, as well as their nature, will be prompted in part by an attempt to achieve some of the advantages of specialization. Since we have already considered these advantages in detail in previous chapters we need not go into the matter here, except to recall that it usually involves the bundling or clustering of closely related activities for a period of time, to the exclusion of others during that period.

Some activities involve decision-making rather than performance. In an attempt to specialize, the individual might isolate some related decision problems for a specific time of the day or week, and performance might follow at a later time. In a sense the decision-making activities are likely in the long run to be more significant than the performance activities for the results of decision-making activities will control and influence the nature of the performance activities undertaken.

From the viewpoint of economics we are primarily interested in decision-making, which we may look upon as one of the unique aspects of human and social behavior, as against performance,[4] which is a much more mechanical matter. This is especially likely to be true when we consider the relations between individuals. As a consequence we shall develop a concept of role that can be employed in such a way as to emphasize the decision-making aspect.

Every voluntary act requires that a decision precede it. At the very least a person must decide whether or not he wishes to carry out the act. Thus, we can visualize all activities as being made up of two components: one involving decision-making, and the second involving attempts at performance in accordance with the decision. We shall look upon the concept of a role, and the activities that go on within a role, in accordance with the dichotomy we have just mentioned. But before continuing this line of thought, let us proceed with our *Gedankenexperimente* and see how roles might arise within a group.

Individuals can conceivably live and work in complete isolation from each other, but if they did so they would hardly be taking advantage of the possibilities inherent in group existence. The fact of the group permits some individuals to specialize in some things while others specialize in other things, and allows them to exchange the results of their specialization to the mutual benefit of both parties. In other words, simple considerations of efficiency resulting from the trading of services or goods may lead to the establishment of roles. This is the application of the old comparative-advantage idea. However, this specialization between individuals rather than within an individual at different points in time does bring a new element into the picture—the element of *mutual recognition.* In order for the specialization of activities between individuals to be carried out effectively, and for the type of specialization to have some degree of durability, individuals must have a mutual awareness of each other's activities. If we began with a group of isolated individuals communicating with one another, we would expect a growing awareness that some people perform certain activities much better than they perform other activities, but that the ordering of proficiency differs considerably for different individuals. We might further expect that a simple exchange of services on a one-time-only basis would be arranged between some individuals. Where the exchanges were successful it would be reasonable to expect that such arrangements would be renewed time after time and eventually would take on the form of a continuing arrangement. Finally, there would be a gradual recognition on everyone's side of the different parts that various individuals seemed to be playing in the production process. This would be the beginning of role-taking.

---

[4] Performance is more in the domain of technology than in that of social science.

Clearly, role-playing in this sense involves not only the specialization of activities but also a web of awareness or recognition of this type of specialization. If mutual recognition is lacking, then individual specialization of activities would be quite pointless since we cannot expect exchanges to take place without this sense of mutual awareness. But it is important to observe that the mutual recognition need not be with respect to specific individuals. From A's point of view it is sufficient that there be someone, but not any specific person, carrying out a certain range of activities with whom he can indirectly coördinate his own activities. In other words, for an individual to live in a society in which there is considerable division of labor, it is necessary for him to know that various people perform and stand ready to perform various tasks that will help him meet his wants. For a differentiated society operating under principles involving the division of labor to persist, people have to play roles, but it is clearly not necessary that any specific individuals play any specific roles.

For some activities merely individual efforts are not enough. Teams, or groups of coördinated individuals, must be employed. Here a slightly new element enters the situation. If a group of individuals have to perform a certain task, then the activities of the individual members of the group must be coördinated. This implies a possible division of labor between the *coördination* of activities and the actual performance of the team tasks. Many tasks are beyond the strength and capacities of a single individual; many others it would be uneconomic for one person to undertake. For example, the building of a bridge—indeed, the construction of a large variety of structures—is not the sort of thing that can be done by an individual alone.

Let us look for a moment at the reverse side of the coin. Suppose we begin with the notion of a coördinated group of individuals (a team). We perceive immediately that there are possibilities of specializing within the team. Here the problem of coördination and of arranging for different people to do different things becomes paramount, and there is a separation between what we may call performance activities and administrative activities.

Now a team may be formed for a one-time-only task. But the task may be of a kind that could be performed repeatedly to the advantage of various people and under various circumstances. Once this insight is gained by some members of the team they may see an advantage in holding the group together to stand ready to do the same task, or a close variation of it, or closely related tasks, in the future. This would require that the team persist through time. Steps would have to be taken to guarantee the durability of the team, and to guarantee the recognition of the team's activities by the outside world. In other words, performance recognition and durability recognition would have to be obtained. When these steps are taken successfully we no longer have simply a team on a one-time-only basis, but an organization.[5]

All of this is not to suggest that actual roles are developed in precisely the same way as they turned out in our *Gedankenexperimente*. The forces of tradition and history are usually significant in the creation of many real roles. But there is at the very least an economic rationale for the existence of roles and role-playing. If we start with a situation where for some reason roles and role-playing do not exist, then we can readily conceive of the economic desirability of creating them. The economics of specialization suggests that people are likely to be more efficient if they take roles and interchange services than if everybody tries to do everything for himself.

---

[5] We are employing the word "organization" loosely here. See Chapter 10 for an exact definition.

But the argument from specialization and the advantages from the exchange of services are not necessarily sufficient to call for the establishment of roles and role-playing as we are using these terms. We can have an exchange of services without role creation. The desirability of role-taking must lie in part in the advantages of the durability aspect of roles.

Before proceeding to our formal definition of a role, we must say something about the distinction between occupations and roles. It may appear at first blush that our analysis could have been carried out just as well in terms of occupations. If we wish to restrict our analysis to the economic sphere, using this term in the narrow sense, then this is indeed true. However, the concept of an occupation has connotations that are really too narrow for our purposes, especially if we wish to emphasize organizational rather than performance activities. For example, there may be no such *specific* occupation as the vice-presidency of a firm, but we may conceive of a person with that title playing a certain role within the firm. Also, for some analytical purposes it is important to be able to connect economic roles with noneconomic roles—for example, the roles within a firm and the roles a person may play within his family. From this point of view we might well define our basic units of analysis in such a way that they can be employed in both an economic and a noneconomic context.

TOWARD A DEFINITION OF ROLE

To get at the definition of role it will help if we break up the problem into two parts: First, we shall define the notion of a *role interpretation,* and second, we shall use this concept to define the concept of a role.

Before proceeding let us summarize briefly some of the characteristics that we have ascribed to a role. First, role-taking involves the specialization of activities, and the concentration of activities into convenient activity bundles. Second, role-taking implies a repetition of performance. Third, the existence of a role and role-taking implies the recognition of this role by those with whom the role-taker will come into contact in his capacities as a role-taker. Fourth, the existence of a role implies the existence of means to maintain the possible continuity of the role and the possible continuity of all of the characteristics of a role that we have just mentioned.

1. *Role interpretations.* We can observe the *playing* of a role by some individual, but not the role itself. Further, we rarely observe the playing of an entire role, but simply one interpretation of it. In other words, we visualize a role as an abstract entity reflected through its interpretations. This implies that many roles, certainly the more interesting ones, will have more than one possible interpretation. We must, however, be careful to distinguish between a specific play of a role and an interpretation.

The basic component of an interpretation we will call the field of alternatives, or the field of action. These concepts are interchangeable. To begin with, the role-taker faces at least one set of alternatives. He has to make a choice. Once he chooses an alternative he has to perform in accordance with his decision. We visualize this activity—choosing and performing accordingly—as one *move* in the playing of a role. This move is one way of playing within the field of action. If there are 10 mutually exclusive alternatives in our alternative set, and if each alternative has an associated activity, then there are 10 possible moves.

Now there often will be more than one field of action in a role interpretation. That is, the role-taker may have to face more than one of the alternatives in a unit period. Each set of alternatives of the same kind we conceive of as the same field of action.

In describing a role interpretation something must be said about the relative frequency with which each field of action comes up. Suppose our two fields of action are A and B; then a role interpretation in which these two fields come up equally frequently is different from one in which A comes up, say, 10 times to each time that B comes up.

A complete description of a role interpretation would consist of a statement of all the fields of action involved in that interpretation, and a statement of the relative frequency with which each field of action comes up. Thus, (2A, 3B) may be a symbolic way of writing a role interpretation in which A is one field of action and B another, and in which the Arabic numerals indicate the relative frequencies with which they come up. Implicit in such a description is the use of a suitable time period so that we do not become involved in fractional roles.

An additional point to note is that our definition of a role interpretation implies no specific sequence in which the fields of action have to be played. Of course, there may be situations in which the necessary activities are sequential. In that case, the problem of describing an interpretation of a role becomes more complex. One possible way of handling such situations is to consider each sequence of activities as one of the alternatives within our field of action. Where this can be done the problem is reduced to fields of action that can be treated in a nonsequential manner. For the time being let us assume that this is always the case for the problems we will have in mind so that we may the more easily proceed with our definition of a role.

2. *Some complex role interpretations.* We now have to consider briefly some complicated types of role interpretation whose significance will come out more clearly later.

In some situations the role-taker may expect the fields of action to come up with specific frequencies, but in others they may be contingent on circumstances beyond the control of the role-taker. There may, nevertheless, be a specific probability that suggests the frequency with which a field of action can be expected to come up on the average. The fields of action become effective, so to speak, when the role-taker observes the consequences of more routine decisions, or as a consequence of either internal events or outside happenings.

To characterize a role interpretation containing a contingent field of action we may associate a number with it that represents the mathematical expectation of the frequency with which that field of action will have to be faced within the unit period. Of course, in any specific period the frequency with which the contingent field of action comes up may be greater or less than the mathematical expectation, but the average frequency should equal the mathematical expectation. If a role interpretation does not contain any contingent fields of action, or any of the fields of action described below as indefinite, then we will refer to it as a fixed role interpretation.

In the early stages of undertaking a role, a player may not know precisely what a possible role interpretation is. There may be an exploratory period in which he cannot completely characterize the interpretation that he wants to give to the role. Obviously, not all role interpretations can be well defined at the outset. They may become much better defined as a result of experience. However, some roles, especially at the highest executive level, may never become completely definite. In other words, there is a limit to the extent to which it is possible to routinize decision situations. Some degree of latitude ought to be left in order to handle situations which are not completely foreseeable but come into the purview of the role on the basis of a vague description or vague knowledge of some of the responsibilities or duties attached to the role. We shall refer to such vague responsibilities or duties as *indefinite* fields of action.

Another way of looking at the same situation is to say that there are some contingent fields of action to which one cannot attach a mathematical expectation of its probable frequency. In still other cases the exact nature of the contingency may be incompletely known. We shall refer to role interpretations involving such instances as incompletely characterized role interpretations. Unless otherwise stated we shall assume in our analysis that the role interpretations with which we are concerned are completely characterized.

A role interpretation that contains a contingent field of action whose frequency can be described by mathematical expectation is a completely characterized role interpretation. For example, a salesman's role interpretation may be well defined even though he cannot foretell how many customers he will have to handle in any given day. A vice-president in charge of the development of new markets, however, will probably not have a completely characterized role interpretation.

Incompletely characterized role interpretations will often imply the necessity of exploratory behavior. Consider, for example, a set of alternatives that can be described by a "tree" (see Figure 20).

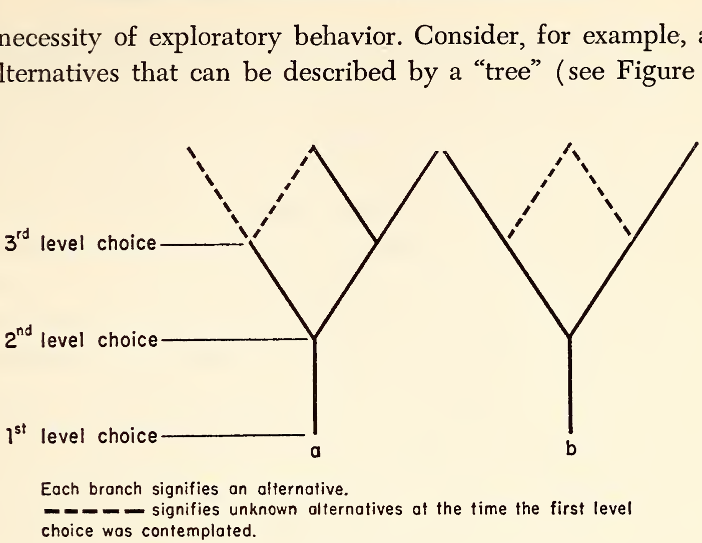

3rd level choice
2nd level choice
1st level choice
a
b
Each branch signifies an alternative.
▬ ▬ ▬ ▬ signifies unknown alternatives at the time the first level choice was contemplated.

Figure 20.

Here a choice among the first-level alternatives opens up a new set of alternatives, which in turn may open up a third set of alternatives, and so on. Where all the alternative possible paths are known we have a situation that is equivalent to a fixed field of action. Each alternative path is really a complex alternative. But some of the alternatives that may open up on the second, third, or higher levels of the "tree" may not be known. This is the situation usually involved with respect to exploratory behavior.

We must distinguish, however, between a movement from a less completely characterized role interpretation to a more completely characterized one, and a shift by a player from one fixed role interpretation to another within a given role. Also, it should not be assumed that all role interpretations are stable. Those containing indefinite fields of action are especially likely to shift to some extent from one period to another depending on the results of explorations and experience.

We have not said anything specific on how role interpretations arise. This can best be done later when we consider roles in the context of organizational behavior. Suffice it to say at present that it will depend on the role-taker's perception of the situation, and on the information communicated to him, information that will suggest to the role-taker some of the fields of action within the role. Also the role-taker will often be informed of rules and prohibitions that indicate some of the fields of action definitely outside the role.

3. *The role concept.* A particular role interpretation need not be the only manifestation of a role. Suppose it is left up to the individual to interpret a specific role that he enters. The role may be so proscribed by the nature of its connection with other roles that only one interpretation is possible. But for a great number of roles that come to mind there is clearly more than one interpretation possible. For example, different people who take on the presidency of a firm may have different notions as to how to distribute their time between the various problems and activities. They may be said to interpret the role in different ways. The role itself may therefore be conceived of as the set of all permissible interpretations. We shall mean by a role simply the set of role interpretations plus some criteria to indicate or suggest which interpretations belong to the set and which do not. We may distinguish between player-recognized roles, public roles, and organizationally recognized roles. By a player-recognized role we will have in mind the set of role interpretations that belong to the role as seen by a specific player. By a public role we will have in mind the set of interpretations that is recognized or at least recognizable by whatever part of the public is influenced by or can influence the playing of the role in some way. Here the nature of the role is perceived by the relevant portion of the public. By an organizationally recognized role we have in mind the set of role interpretations that is recognized or recognizable by members of the organization as belonging to it. It is important to observe our use of the word "recognizable" in this context. Often many members of the organization do not have a clear-cut notion as to how the role should be interpreted although they may have certain ideas as to what they would consider an improper interpretation of the role.[6] Thus, it is not really necessary that the role be

---

[6] Often the role-taker will be told specifically some of the decision and

recognized specifically from an a priori point of view, so to speak, but only that the people involved can say whether a certain interpretation belongs to the role when they are specifically faced with that interpretation.

When more than one person is involved in the recognition of a role, how can the role be described if different people have different visions as to what interpretations belong to the role or do not belong to it? There will always be a portion of an organization that may be said to control the role—to be in a position either to choose the role-player or to reward or penalize him. Now let us suppose that the controlling individuals have varying notions as to how the role should be interpreted. Here are two possibilities: (1) there may be a subset of interpretations that are common to all these individuals, or (2) there may be no such subset. If there is a subset, then we may say that this is the *identically recognized* part of the role. If there is no subset, then we will say that this is a nonidentically recognized role. Clearly, in some cases the former will hold and in others the latter. The importance of these concepts will become clearer in the chapters that follow, where we consider the stability of the role and the interrelations between roles. If all individuals that are somehow connected with a role visualize that role in terms of exactly the same set of interpretations, then we will call that a well-recognized role. In what follows the context will usually indicate how we view a role, but in the absence of any such indication it should be assumed that we have in mind a well-recognized role. Although the concept of a well-recognized role is an extreme simplification of reality, it is useful for analytical purposes. Very few roles are likely in a strict sense to be well recognized, but in the early stages of analysis it is desirable to assume that they are. This is similar to the sense in which such concepts as perfect competition and homogeneous product are employed in economic theory.

---

action areas that are outside the role, or that are prohibited to those taking that particular role.

10

Toward the Definition of an Organization

INTRODUCTION

Given the concept of role developed in the last chapter we may, as a first approximation, look upon an organization as a collection of connected roles that are currently being filled. But the crucial elements, to be considered in this chapter, are the types of roles included and the relations between them that have to exist for a collection of filled roles to be considered an organization. We can, of course, define our terms any way we please but it would certainly be advantageous to work toward a definition that includes most of the entities we would, in everyday parlance, call organizations.

In looking at an organization in terms of a collection of roles we have in mind that the roles are filled. If they are not filled, then we may have an empty organization or a paper organization, but not a real one. Why should we look at an organization as a collection of roles at all; why not simply as a collection of people who are related to each other in certain ways? The answer is, in part, that that would be looking at the problem from too specific a viewpoint. The "collection of people" approach would permit us to look at an organization at an *instant* of time but would not permit us to look at it as a durable entity—sometimes an indefinitely durable entity, when the people in it change entirely. We must define our concept sufficiently broadly so that we need not be concerned with the specific people composing the organization (i.e., filling the roles) at any one time. The labor contract being what it is, people are usually free to belong to or leave most organizations whenever they please, or at least after some prescribed interval (e.g., the army). Thus, we want to look upon the specific members of an organization as replaceable by others outside the organization without necessarily destroying the organization itself. In other words, we conceive of an organization as being composed of replaceable human parts.

But an organization, as that term is normally used, cannot be simply *any* collection of roles, or even any related collection of roles. The roles of a husband and of a breadwinner (e.g., a salesman) may be related in some sense, but they are not necessarily parts of the same organization. Thus, we have to select the types of relationships that can enable us to tell when any two roles are in the same organization and when they are not. Such a selection would give us a criterion for determining which roles, in any specific instance, are within an organization and which are outside it.

We shall approach the organization concept in easy stages. First we shall classify roles into the role types commonly found in every human organization. Each role type will imply a relation or set of relations with other role types. From these relations we shall select those that are crucial to any organization and those that determine whether a given role is or is not part of the organization. There is a close connection between the types of activities that usually go on within any organization and the role types involved in carrying out activities within the activity categories. This will be employed in determining role types and the relations between roles.

Another aspect of the picture to be kept in mind is that specific roles and role relations do not necessarily exist at any point in time.

They may have to be created. Furthermore, they can be created at the same time that other activities within the organization are taking place. Organizations are not quite like ordinary machines, although the machine analogy is sometimes found in the literature. We may appropriately digress somewhat and look at the machine analogy before proceeding toward our attempt to develop a concept of an organization.

ORGANIZATION AND THE MACHINE ANALOGY

It is not uncommon to think of organizations as machines with human parts. While this concept may have some value, it could be quite misleading. If an organization is to be looked upon as a machine, then it must be as a unique type of machine with characteristics quite different from other mechanical devices such as watches, automobiles, engines of various sorts, and so on. Even the analogy with servo-mechanisms of the usual kind, such as a thermostat, does not go far enough. Indeed the interesting aspects of an organization are precisely those in which it differs from the ordinary machine.

Organizations, like machines, may have to be designed with a certain purpose or set of purposes in mind. Suppose, to begin with, that the organization is to perform a given job. That is, there is a given output and it could repeat the output indefinitely, period after period. (Later we consider the possibility of changes in the output of the organization.) From the point of view of the machine analogy, how should the organization be set up to produce the output?

The output implies or determines the scale of operations and the size of the organization. Specifically, we have to find a set of activities that will produce the necessary output. Simultaneously we have to determine the factors of production, that is, the combination of labor, equipment, and materials, that could carry out the requisite activities. In the usual case, where there is more than one set of factors of production that could perform the activities and more than one set of activities that will produce the output, we would pick an activity bundle and an associated factor bundle in accordance with some criterion. For example, we usually would want to choose that factor bundle and its related activity set that would cost least.[1] We associate its costs with each bundle of factors. Where any bundle can carry out more than one set of activities to produce the output, then we associate with that bundle the least costly set of activities. We then rank the bundles of factors in terms of their cost and pick the least costly factor bundle.

The task of organization here is purely mechanical. We have to obtain a set of activities that dovetail with each other in terms of both the sequence and the nature of the activities. In other words, for each person in the organization we set up performance specifications for the activities so that activities dovetail; at the same time we set up a sequence of activities and a spatial arrangement so that they all fit in with each other. This done, we presumably have a machine that will produce the desired output.

If we want an organizational machine that will produce variable outputs, the situation is not fundamentally different. Our task is to choose factor bundles that can carry out more than one set of activities on order. If the set of factors is chosen on the same basis as before, our machine will then produce a range of outputs. Of course, the factor sets would have to be so chosen that part of the activities possible included those minimal communication activities necessary to pass on the orders required to make the machine shift from the production of one output to the production of another.

As a matter of fact, almost no organizations of any size are of the kind just described. The employment of human parts as well as some other elements in the situation makes the strictly mechanical analogy inappropriate. Two critical elements have to do with information and with the fact that the organization itself (rather than some outside agent) will at least in part determine its objective and outputs. Since many elements in the situation are not known, an organization constructed along the lines of a simple machine will usually not work.

---

[1] This view of the problem is referred to as the "classical" organization theory approach in James G. March and Herbert A. Simon, *Organizations,* John Wiley and Sons, New York, 1958, chap. 2. See also J. A. C. Brown, *The Social Psychology of Industry,* Pelican Books, Baltimore, 1954, for a summary of the social psychologists' view of the problem—a view that is antipathetic to the classical approach.

Consider some of the elements which, in fact, are not known but which are presumed to be known in the machine analogy. (1) The required output may be changing, but the nature of the output and its amount have to be discovered by the organization itself. (2) The activities to be performed may be changing and also not completely known. (3) Similarly, the capacities of the factors of production may be changing; in any event they are likely to be incompletely known. (4) Finally, the availability of factors to the firm for purposes of replacement are also incompletely known. In sum, we have to build our organization in such a way that *simultaneously* with performance we have the means of gathering information about the changing environment and about the possibly changing capacities and activities within the organization.[2] The operations of the organization will yield not only output but also information about activities, factor capacities, and so on. Thus, the organization has to be constructed so that its human parts can adjust to one another in light of changing information and do this at the same time that they are performing in other ways. Indeed, for this reason it is often desirable to have more than one role interpretation within a role. A role-player can then attempt to utilize the information obtained in the course of role-playing in order to shift from an inferior role interpretation to a superior one.[3]

Also, since the organization itself determines in part its specific objectives, we have to look at the organization not as the usual kind of a machine but as an entity in which the user of the "mechanism" and the "mechanism" itself are part of the same entity. With such considerations in mind, we shall highlight in what follows the relations of communication and authority between roles, although many of the roles themselves are neither unique nor fixed for all time.

---

[2] Here we might mention the connection between the values and motivations of members of the organization and performance—especially the connection between these elements and the nature and operation of the organization. Cf. March and Simon, *op. cit.*, chap. 3. We shall touch on these matters later on in this book.

[3] I.e., from the player's point of view. From the organization's viewpoint the shift is sometimes in the opposite direction. This is part of the problem of organizational pathology that we cannot go into at present.

ORGANIZATION ACTIVITIES AND ROLE TYPES

The problem of identifying and classifying the various types of roles is rather difficult because we are immediately faced with a large variety of possibilities. It would seem desirable to limit our classifications to a sufficiently small number so that we are not overwhelmed, and at the same time to have a sufficiently large number so that we can make all of the distinctions useful in analysis. From this point of view we shall attempt to classify roles in connection with the usually essential activities of any organization. But first: What is there about a role that would determine its classification?

The essential components of a role are its fields of action. The well-defined role may be characterized by one or more of its fields of action. But, as we have defined the concept, a role may have more than one interpretation. Usually, however, some fields of action will be common to all interpretations of the role and it is these fields of action, or a member of this subset, whose characteristics will determine the nature of the role.

The main point is that fields of action are what we really try to classify. If a field of action is in a certain role interpretation, then we can apply the same category to the role interpretation. That is, a role interpretation may be said to be in category A if it contains a field of action in category A. Of course, a role interpretation may contain fields of action in several categories. It is important to note that for our purposes we need not have exclusive classifications. We are trying to arrive at relations between role-fillers, and many types of relations may exist between one role-filler and another within an organization. In other words, we seek means of classifying role interpretations that are suggestive of the relations between role interpretations in different roles. We classify the roles in a similar way. If a role contains a role interpretation that falls into category A, then we shall designate the role as belonging to category A.

Another way of looking at the matter is that we seek shorthand means of talking about roles. It would be too cumbersome to speak of "a role that contains a role interpretation that contains a field of action in category A," and hence we shall simply categorize roles in terms of the fields of action they contain. All this should become clearer as we proceed and become more specific.

Observe, however, that a role, since it may contain many fields of action, may fall into more than one category. Broadly put, some people may play very simple roles with only a single characterization, but others may play rather complex ones.

Some roles will contain contingent fields of action—fields of action that are well defined but are not acted upon in any routine or periodic manner; they are considered only in response to certain external circumstances, or in response to the observation of the consequences of playing some of the routine fields of action. In other words, a role may be so specialized that a player's expectation in that role is to ready himself for certain contingencies, although he may occupy himself with other activities when these contingencies do not arise. But such roles can be categorized in the same manner as other roles.

Often some fields of action dominate a role. Let us look at it first from the point of view of those members of the public or of the organization who have to recognize the role. They may conceive of its having a variety of interpretations, but at the same time they may expect that in all interpretations at least certain types of activities will be performed. For example, there may be various ways of playing the role of the university professor, but at a minimum the members of the university community will expect the professor to meet students at least for certain periods during the school term. Similarly, in other roles such minimum expectations usually exist. Thus we might say that, since by their very nature these minimum expectations involve fields of action common to all role interpretations as seen from the point of view of the organization, these are essentially the dominant fields of action within the role. When we speak of specific role types without qualification, then, we usually have in mind the role type in terms of its dominant fields of action.

Since we are interested in how roles are related to the concept and the operations of an organization, let us try to categorize roles in close connection with the major activities (operations) usually necessary in any organization. A discussion of organizational activities and related role types follows.

1. *Direction activities and direction roles.* In every organization someone, or some group, has to determine what is to be done, to communicate decisions to subordinates, to anticipate performance by subordinates, to observe and examine the results and consequences of his communications, and finally, to make new decisions for future performance. Obviously, the direction activities will be closely related to other activities, especially those of linkage and performance. The direction roles follow immediately from the direction activities of the organization. Those people in roles that contain fields of action that in turn contain "decision-activity" alternatives with respect to the determination and direction of specific performance may be said to be in the direction roles.

2. *Linkage activities and linkage roles.* At the outset we may subdivide this category into internal linkage activities and external linkage activities. By linkage activities we mean those activities involved in linking or connecting the decisions and activities of one individual with those of other individuals. The decisions and activities linked may be entirely within the organization, in which case we call them internal linkage activities, connected with the outside world, when we call them external linkage activities.

The external linkage activities, as their name suggests, are the activities that maintain and determine the relationship of the organization and the outside world. At the very least this involves the receipt of information from the outside world and in most cases the recruitment of new members of the organization. At the other end external linkage activities will involve arrangements to render services or sell goods to the outside world, or parts thereof, and the transmitting of information for these and other purposes. By linkage activities we shall have in mind those that usually come under the headings of communication and transportation. Individuals who are in roles involving such decisions and activities will be said to be in linkage roles.

3. *Performance activities and roles.* Performance activities are whatever the organization is primarily expected to do in its relation to the outside world. For example, the performance activities in a firm are those activities directly connected with the creation of goods and services. The average worker on an assembly line in an automobile plant is engaged in performance activities and is in a performance role. Similarly, noncommercial organizations usually have something they are expected to do vis-à-vis their relations with the outside world. When the members of the French Academy, for example, are engaged in working on their French Dictionary they are carrying out the performance activities of that organization. Of course, in this particular case it may be debatable whether the formal mandate of the organization with respect to performance is the real mandate. Perhaps it is their main task to choose French immortals in order to glorify the intellectual contributions of France. Whatever the interpretation, there are some activities connected with performance directly and others not so connected, or related only to the maintenance of the organization.

A clear-cut association exists among direction, linkage, and performance activities and among their related role types. Obviously, the linkage activities connect direction and performance. Basic decisions are made at the direction level and are communicated through the linkage activities to performers, who in turn do the actual work.[4] Reversing the circle, the nature of the performance has to be observed and the pertinent information resulting from such observation needs then to be conveyed through the linkage activities to the directional level, where with these further data additional directive decisions may be made. But this is far from being the entire picture.

4. *Control activities and control roles.* Side by side with the

---

[4] Most classificatory systems are vague at the edges. In practice, how a particular activity is to be classified is a matter of judgment and will depend on the specific nature of the problem. Some transportation activities may often be usefully classified as performance activities, e.g., the transportation of materials from one point on an assembly line to another.

direction-linkage-performance mechanism we have a control-linkage-performance mechanism. The reason for the existence of the latter complex of activities and roles lies in the belief that the former complex will not work smoothly or as smoothly as desired without the existence of the latter. That is to say, directives may not be accurately conveyed, or if accurately conveyed performance may for some reason not be in accordance with the directives, or decisions may not be made in the light of performance. In other words, the links between various types of activities may not operate as they should, or individual performers may not operate either in accordance with expectations or in accordance with their relations to other performers. For some reason or other, things do not always work out, if left to themselves, as some members of the organization would like to see them work out.

Another way of looking at the matter is that mistakes may be made which have to be corrected, or some degree of free choice leads to activities that are considered undesirable, and this degree of choice may have to be limited. Somehow we have to make sure that activities are in fact coördinated in accordance with some concept of common aims. As a consequence, some degree of control is usually necessary. In an organization with specialized roles, there will be some person or group of people for whom control activities and control roles will comprise at least a part of the activities within their fields of action.

The control activities involve assessment of performance, comparisons with formal or informal performance standards, and sometimes decisions with respect to rewards and penalties as a result of such comparisons. Of course, any decisions taken have to be communicated to the relevant people and acted upon. They obviously involve the linkage activities.

Enough has been considered to suggest that the relations between different types of activities within an organization are not simply hierarchical or unidirectional. The pattern is much more complex. Direction and control have to be performed simultaneously, and means have to exist for the coördination of these two types of activities. In other words, we may conceive of a direction-control complex which is closely interrelated in the sense that information of each type of activity has to be relayed to the other activity center, and both are connected with the linkage-performance mechanism.

5. *The payoff-distributing activities and payoff roles.* The activities of the organization lead to consequences some of which have their impact outside the organization, others within the organization. For example, the income of a firm is the consequence of its activities whose impact is in part within the organization. Some means has to be found to decide on the distribution of a firm's income among its personnel, stockholders, or others. We may look upon such decisions as determining the allocation of the payoff of the organization and those engaged in making the decisions as being in payoff-determining roles. To some extent the distribution of the payoff may be used as part of the system of control. For example, a bonus system out of the profits used as an incentive to personnel would be part of the control system. Similarly, penalties and fines increase firm income and are indirectly a means of redistributing firm income from some people to others.

6. *Role-determining activities and roles.* The roles of the organization are not given, so to speak, from the outside. They have to be determined by the organization. Various members of the organization will be in a position to determine or change the nature of the roles within that organization. The people who do so may be said to be playing, in part, role-determining roles. In some cases the roles may appear to be established in small part by a document or a charter or something of that nature. However, there are usually means of changing such documents or charters, and those people who can change them are also in role-determining roles. These are crucial roles in any organization.

7. *Player-determining activities and roles.* Once the roles are determined, the players to fill them must be chosen. Anybody who can choose another may be said to be in a player-determining role. This is almost a self-evident category and we need not consider it in detail.

8. *Advisory activities and roles.* All of the roles that we have considered involve directly the operations of the organization. Some

members of an organization may be engaged in the analysis and evaluation of information. They do not themselves make organization decisions. However, the fruits of their efforts are often taken into account by those who do make such decisions. For the sake of completeness we include this category in our list.

RELATIONS BETWEEN ROLES

In any large organization a complex web of relations usually exists between roles. These relationships will manifest themselves through the fields of action and their consequences contained in the various role interpretations.

Now, relations between roles can exist at any of the three levels of the decision-activity-consequence sequence. First, there is the decision-making segment; second, there is the actual performance or carrying out of decisions; and third, there are the consequences of the decisions and their performance. The relations between roles may exist at the decision level, at the performance level, at the consequence level, or at some combination of these. Some relations between roles may manifest themselves in the day-to-day operations of the organization, while others may be potential and contingent on nonroutine circumstances.

Of course, we cannot go into detail on the entire web of possible relations. But some types, such as the authority, influence, and communication relations, are crucial to the functioning of an organization. We shall examine them in detail in a later chapter. For present purposes we are interested in commenting on relations only to the extent necessary to elucidate our concept of an organization, since we view an organization as a set of related roles. Roles operating in isolation, so to speak, are not part of an organization. A role in an organization has to be related to some other role within the organization. But roles within an organization are also related to roles outside an organization—for example, those of the buyer of one firm and the sellers of another. Thus, the mere fact that roles are related to each other is hardly enough to determine the boundaries of an organization. But since the relatedness of roles is a necessary condition for a set of roles to be an organization, we suggest here at least some ways in which roles may be related to each other.

What do we mean when we say that two roles, or more precisely two fields of action in different roles, are related to each other? We mean that the choice made in one field of action is somehow affected by that in the other. This may come about in a variety of ways. The alternative in one role may be narrowed or widened by the choice made in the other. In the case of the complete authority relation the alternatives open to the one receiving an order may be so narrowed as to leave only one alternative and hence no choice whatsoever. In other cases the relation may be indirect, so that the choice made in one field of action sets in motion a train of events that alters the alternatives or rule of choice open to someone in another role.

Now the names we have given to various role types suggest some of the relations in which they may be involved, but certainly not all the relations. The communication relation will usually exist between a large number of roles. Between roles within the same category a coöperative relation may exist at times, but at other times it may be competitive.

Individuals playing certain roles may be expected to work together in performing some activity, although these individuals do not, in their decision-making, impinge upon each other. For example, a group of men erecting some structure may be said to be coöperating, and their role relationship is obviously a coöperative one. The work of a committee involves coöperative role relationships at the decision-making level. On the other hand, a team of men who are, say, part of a road gang involves such relationships primarily at the performance level. The relations are symmetric: if A cooperates with B, then B coöperates with A.

But the most important relations in an organization (such as authority) will usually be asymmetric or unidirectional. Some examples may be the best way of indicating what we have in mind. If X gives a command that Y has to carry out, the relation is an asymmetric or unidirectional one. X is at the giving end while Y is at the receiving end. Y is not in the same relation to X as X is to Y. At the decision level, X may make decisions that Y has to take into account in making his decisions, but not vice versa. At the decision-performance nexus X may make a decision that Y has to carry out. Similarly, at the performance level X may do something before Y can proceed with his activity. This sort of thing can occur on an assembly line. X may have to provide the bolt before Y can screw it on. Obviously, this is a very wide and significant category of relations in an organization.

A specific complex relationship may be a combination of simpler relations. Consider the stockholders and the board of directors of the corporation. Clearly no single simple relation indicates their "relations." The stockholders operate as a group in electing the board of directors. Hence, vis-à-vis each other they are in a coöperative relationship, but vis-à-vis the board of directors they are in an asymmetrical relation. At the same time, since the board of directors will decide on the dividend rates, the directors are in a coöperative relationship with each other and an asymmetrical relationship with the stockholders.

Finally, let us consider briefly some of the main role types in sets of two and examine some of their relations. For example, direction and performance clearly suggest a unidirectional relation. Someone directs and someone else acts in accordance with the direction. Control and performance suggest another unidirectional relation. The relation between someone in a player-determining role and a potential player is of a similar nature. By looking through the list of role types mentioned in the last section and taking them two at a time, we find other combinations.

Enough has been said to suggest what a number of the strands in the web of relations within an organization may be like, and to give meaning to the notion that an organization is a set of related roles, although, as we shall argue in the next section, it is not *any* set of related roles.

Contingent relations. This is not really a new type of relationship but may take on the form of any of those considered above. Some roles contain contingent fields of action, that is to say, fields of action that are acted within in response either to certain external events (i.e., external to the organization) or to certain internal events. A contingent relationship will become effective, so to speak, when the role-taker observes certain consequences of the more routine fields of action, or as a consequence of external events. The distinguishing aspect here is not the nature of the relation, but the fact that it does not exist routinely nor is it likely to manifest itself in the routine activities of the organization. Nevertheless, not only must the existence of contingent relations be recognized, but these may be the most important from the point of view of the functioning of the organization, since they are the ones that will manifest themselves in transitional and emergency situations.

THE CONCEPT OF AN ORGANIZATION

Roles that are related to each other may be spoken of as role complexes or role structures. By a *role complex* we mean *any* collection of related roles. The slightest reflection should indicate that there are a great many types of role complexes possible. The eight role types set out above can be combined in any grouping that we please. However, some role combinations have certain characteristics that we may want to distinguish—especially from any random collection of related roles. These would be the characteristics that would make the particular role complex especially effective in some respect.

Three role types can usually be combined in a useful way: the direction roles, the linkage roles, and the performance roles. Since any role complex containing these interrelated roles can do work, we shall look upon such a complex as a working group. In other words, for work to be performed (i.e., for purposeful things to be done) it is necessary to have direction, and therefore someone in the directing role or roles; it is also necessary to have at least the communication linkages between direction and performance and hence some individuals must be in linkage roles; and finally, some people must actually do the work and hence be in performance roles. But, we shall see, if such a group does not contain other role types, then its activities are limited, and it certainly cannot be looked upon as an organization as that word is ordinarily used.

One characteristic of the working group is its dependence on some other entity for its creation and continuity. For example, since there are no player-determining roles in the working group, the group is dependent on an outside entity for the replacement of members of the group who leave. If we add to the working-group role complex a player-determining role, we have a somewhat more independent structure. To facilitate discussion let us refer to such a role structure as a "team." Similarly, we shall refer to any role structure containing a working group plus some other related role types, but not all of the role types that are required to make up an organization, as a suborganization.

Examples of "teams" are athletic teams. Their role may be prescribed in advance by the nature of the game and by the league to which they belong, but they may have the option, as they usually do, of choosing their own players. While this is not strictly the case with some athletic teams, it will do for our purposes in order to distinguish this role complex from that of the working group.[5]

As we add role types to our suborganizations we will usually find that they become more independent of any larger organizational entity. Clearly, a suborganization containing the player-selection and payoff-determining roles will be more independent than one lacking the payoff-determining role. A completely independent role structure would contain all of the role types. An independent and unlimited organization would be such an entity.

We now come to our main concept, that of an organization and its variations. By an *independent organization* we shall mean a role structure that contains a working group plus the complete player-selecting, role-determining, and payoff-determining roles. We will show later in our analysis of an organization that it is quite necessary for most organizations, if they are to be viable, to have all of these role types within it. However, it is *role-determining* roles that are crucial to the organization. In all cases we will view a role structure without the role-determining roles as not being an organization.

---

[5] The reader may wonder why we want to distinguish between different types of suborganizations. There may be many instances where such distinctions are useful, but consider at least this possibility: One of the central problems of organization is suggested by the dichotomy of centralization versus decentralization. To what extent should an organization be centralized, and in what way? One alternative would be to attempt to decentralize in terms of separate working groups, while another would be in terms of separate "teams," or in terms of still other types of suborganizations.

A little reflection clearly indicates the crucial nature of the role-determining roles to an organization, and to its independence. The relation between a role-determining role and other roles in an organization is asymmetrical and is an authority relation. Anyone who controls the roles of an organization controls the nature of the organization, since by changing the roles he can change the nature of the organization. Reducing the roles will diminish the organization or even eliminate it if role reduction goes far enough. Similarly, the successful expansion of an organization depends on what is done to the roles—how new roles are added and other roles are changed.

The role-determining roles may also be said to determine the bounds of an independent organization. Roles that are subject to (i.e., can be changed by decisions of) the role-determining roles are within the organization, whereas roles that cannot be determined by decisions of those in the role-determining roles are outside of the organization.

At this point some remarks are necessary to indicate how we shall view the payoff-determining role. We have used the word "payoff" more or less in the sense of the revenue of the organization. There are two aspects involved: first, the determination of the gross revenue of the organization so that it can carry out its activities; and second, the determination of the distribution of the revenue within the organization. Let us refer to the gross revenue as the aggregate payoff to the organization. If the aggregate payoff is not controlled by an outside agency, we shall view the organization as independent. If it is, we shall regard it as dependent. However, we shall assume that the organization is free to determine the payoff distribution within the organization. Otherwise we shall consider it a restricted or limited organization.

An organization may have limits put upon it in determining the nature of some of its roles. For example, some roles or their interpretation may in part be determined by the legal system. With respect to firms as organizations the payoff distribution is in part legally determined through the system of taxes and other laws. In this sense almost all real organizations are limited to some extent. But for ease in proceeding with our analysis we shall assume that the organizations with which we have to deal are unrestricted organizations.

It is important to note that in the case of the independent organization there is no necessary presumption that the aggregate payoff depends on the performance of the organization. It may or it may not, depending on the nature of the organization.

A DEFINITION OF THE FIRM

We now come to the point where we want to define a special type of independent organization, the firm, in terms of the concept and definitions developed previously.

By a firm we will have in mind an independent organization whose destiny is determined by the magnitude of the aggregate payoff, and in which the aggregate payoff depends directly on its performance and specifically on the production and sale of services or goods. Since the firm is defined as an independent organization, it is automatically conceived of as a set of related roles. Furthermore, all these roles, in one way or another, are related to the payoff. The special characteristic of an independent firm is that its destiny is connected with the magnitude of the payoff. In specific terms the success or failure of the firm is obviously related to its profits and losses—to the magnitude of the payoff, as well as the payoff distribution. For example, if all profits are distributed in dividends, there may be no growth, whereas if this is not the case there may be growth. An additional point is that performance is also related to payoff either directly or indirectly.

From our definition we can distinguish organizational entities that are firms and those that are not. An institution like the Red Cross is clearly not a firm since its existence and destiny depend upon other criteria than the income received in connection with the sale of services or goods. A government is not a firm for the same reasons. However, a hospital may or may not be a firm depending on how it is financed and the nature of its charges. If the hospital finances itself through the charges for services to its patients and through no other means, then it is clearly a firm. If it is subsidized or financed through government grants, it is not a firm although it may be an independent organization. That is to say, in the latter case the grants are not rendered in such a way as to interfere with any of the role types contained in an independent organization. If, on the other hand, the grants were employed and determined in such a manner as to control or direct any of the activities of the organization, then the hospital in question would have to be looked upon as a limited organization.

The reader will observe that there is nothing in our definition that depends on the objectives of the organization. For example, we do not define firms as profit-maximizing entities. Indeed, we want to leave the matter of objectives in the specific sense and objective determinations entirely free, since we will want to use the definitions we have worked out to analyze how objectives are determined, and to what extent firms may or may not be profit-maximizing entities. In contrast to the current conception of economic theory we do not assume that the objective is given at the outset; it is determined through the internal operations of the organization.

11

The Authority-Communication Mechanism

AN ORGANIZATION functions through an authority system. That is, the decisions of the organization are made and carried out by decision-makers and actors distributed at various points within the organization. The decision-makers have different degrees of power in determining the activities of the organization. They exercise it through their authority. It is through the authority system, we might say, that the activities of the organization are set in motion. We must, therefore, examine how decision-making, authority, and organizational activities are related to each other.

Our main aim in this chapter is to introduce the concepts of authority and communication into our organizational schema, and to indicate briefly the nature of the authority-distribution problem. It is left to a later chapter to look at the authority-distribution problem in greater detail.

THE MEANING AND NATURE OF AUTHORITY

As much as possible we will want to tie our major organizational concepts to our key concept of role and its components, the fields of action. Authority may be viewed as a unidirectional relation between two fields of action contained in roles played by different individuals. Suppose that the field of action of one individual contains two alternatives, V and W, and that of the other individual contains the alternatives X and Y. Now if these fields of action are related, then the alternatives are related. If each time the first individual chooses V the second one chooses X, a relation clearly exists between the two fields of action. If in addition when the first individual chooses W the other chooses Y, a functional relation may be said to exist. For each choice of the first individual there is only one specific choice possible for the second one. Now, if in addition the choice of X by the second individual does *not* presuppose any definite choice by the first one, then it is clear that the relation is unidirectional; it flows from the first individual to the second, but not the other way round.

Ordinarily we understand a position of authority to exist when one person gives orders and the other is seen to carry them out. But this is essentially similar to the relation indicated in the preceding paragraph. When the first person gave the order he picked a specific one out of an alternative set of orders that he could have given, and the second person in following the order circumscribed his choices accordingly.

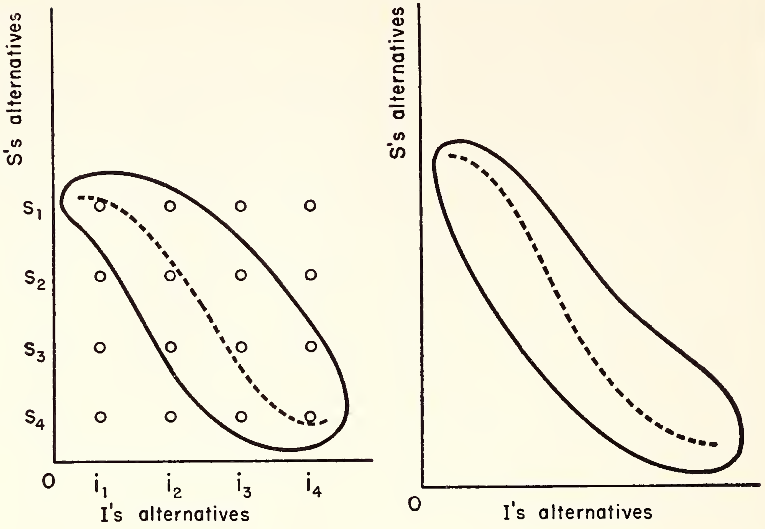
But there is something more to the authority relation. Authority presupposes that the sort of relation just discussed is mutually recognized to exist by the role-players for some segments of their respective roles. In other words, there has to be a prior understanding (or contract) between the potential superior and the potential subordinate that they stand in such a relation, which in turn involves the recognition by the subordinate that the superior has some degree of discretion with respect to his potential orders. We need not concern ourselves at present with how such an understanding is arrived at. It will, of course, lead to certain mutual anticipations by both parties.

Now, where such prior understandings do not exist, and where one person restricts his set of alternatives as a consequence of receiving a communication from another (be it in the form of a command or in terms of declarative or normative statements), we might say that an *influence* relation exists, but not an authority relation. In an organization the authority relation may be said to exist between two roles in the sense that when the role-players undertake their roles they are told or somehow come to understand that such a relation exists between the people filling these specific roles.

We may distinguish between complete authority and qualified authority. Complete authority exists when there is a one-to-one correspondence between the choices of the superior and those of the subordinate. Qualified authority exists when there is, at least in some cases, a one-to-many correspondence between the choices of the superior and of the subordinate. In the case of qualified authority the alternatives of the subordinate are reduced by the actions of the superior, but some element of choice is still left to him.

A diagrammatical illustration may help to clarify some of these ideas. We employ some of the elementary notions of the theory of sets.

In Figure 21A the alternatives of the superior's field of action are shown on the ordinate by s_1, s_2, s_3, and s_4, the alternatives of the subordinates are shown on the abscissa by i_1, i_2, i_3, and i_4, and the combinations of the superior's choices and the subordinate's choices are shown by the dots in the figure. (This is the Cartesian product of the superior's and subordinate's fields of action.) Each dot represents one alternative chosen from the superior's field of action and another chosen from the subordinate's field. The set of dots joined by the broken line in the figure may be said to represent the complete authority set. That is, for each choice from the superior's alternatives there is one and only one choice from the subordinate's alternatives. In common language the superior gives an order and there is one and only one way for the subordinate to carry it out.



A

B

Figure 21.

Figure 21B illustrates the same phenomena, but in this case the alternatives are assumed to be continuous. The entire space within the figure represents all the possible joint choices by S (the superior) and I (the subordinate). The line once again depicts the idea of complete authority, and the enclosed area depicts qualified authority.

In both figures the set of points within the curved figure represents qualified authority. An examination of the curved figure reveals that for some choices of the superior there is more than one choice open to the subordinate. For example, if the superior chooses s_4, then Figure 21A illustrates that the inferior may choose between i_1 and i_2. In sum, we can see that where the relation between the superior's and subordinate's alternative sets is a one-to-many correspondence (a relation that is not a function in the technical sense), only qualified authority exists. Where the relation is a one-to-one correspondence, complete authority exists.

We must make one more distinction before we can relate the concept of authority to organizational activity. We have to distinguish between the superior's perception of the authority relation and the inferior's perception of it. They clearly need not be the same, and in many cases they will not be. In Figure 22 we draw the superior's perception of his authority by a curved figure which is bounded by a continuous line, and the subordinate's perception by the figure bounded by the broken line. If their perceptions were identical, of course, the figures would be on top of each other. Where they are not identical the figures may nevertheless intersect. How they intersect or overlap will determine the area of effective authority, i.e., that segment of alternatives of the superior's field of action within which the choices made by the superior will be responded to in a way satisfactory to him. Effective authority will exist for that portion of the superior's field of action in which the subordinate's perception is contained entirely within the superior's perception. This idea is illustrated in Figure 22. The segment *AB* of the superior's set of alternatives is an area of effective authority

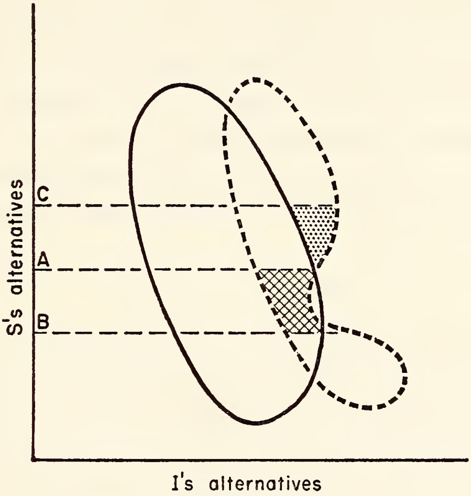


Figure 22.

since all of the responses by the subordinate associated with these alternatives (shown by the cross-lined section in the figure) fall within the superior's perception of what the responses should be. On the other hand, the segment *CA* does not represent an area of completely effective authority because some of the responses that the subordinate would make fall outside the superior's area of perception. (See dotted area in figure.) This last is an area of uncertain authority that may lead to friction and in some cases to eventual correction. The point is that outside of the segment *AB* the subordinate's response choices may not turn out as the superior expected. In the extreme case in which the two perception areas do not intersect at all no effective authority of any kind exists.

COMMUNICATION

The concept of communication is in some ways very similar to that of authority.[1] The communication relationship is also a unidirectional one. Indeed, we can analyze the concept of communication much the way we analyzed the concept of authority.

Authority is a relationship between the decisions of one person and the action responses of another; communication is the relation between messages. Let us call the two parties to the communication relation the "sender" and the "receiver." The sender's field of action is all the alternative messages he can send. The receiver's field of action is the alternative interpretations he can give to the signals he receives. Following the diagrams in Figure 21 we may define precise communication as that relation in which there is a one-to-one correspondence between messages sent and messages received.

---

[1] The mere fact that the communication relation is in some way similar to the authority relation does not mean that it is in fact *always* an authority relation. It may be an authority relation, but it need not be. For example, there may be no understanding on the part of the sender and receiver that the receiver is obligated to receive a message although he may do so voluntarily. In other words, when communication takes place there need not be any necessary sense of *obligation,* understanding, or agreement by both parties to transmit and to interpret "messages," although they may do so voluntarily. Cf. C. Barnard, *The Functions of the Executive,* Harvard University Press, Cambridge, 1956, pp. 172–184.

With respect to imprecise communication there will be, at least in some cases, a one-to-many correspondence. We can use Figure 21 to interpret these ideas. The line (the function) represents precise communication and the area within the curve represents imprecise communication. In the case of imprecise communication, the receiver has the choice of more than one interpretation.[2]

As with the concept of authority, the sender has a perception of the alternative messages and the receiver has a perception of the alternative messages. These two concepts can also be depicted as we depicted the perceptions of the authority relation in Figure 22. The sender's perceptions can be represented by one enclosed area and the receiver's perception by the other. How the perceptions overlap determines the area of effective communication. If the two areas are identical, communication is effective in all instances. In such cases there is no possibility of misunderstanding. Messages sent within the nonintersecting areas may lead to misunderstanding in the sense that the receiver may choose an interpretation not in the sender's perception area, and thus not one of the permissible interpretations that the sender has in mind. The analysis of the overlapping areas is exactly analogous to that of the authority relation illustrated in Figure 22; hence we shall not examine this aspect of the situation in detail here.

Now authority and communication are related in an intimate way. For it is through communication that authority may be said to function. To put it simply, the orders given cannot be put into operation unless they are effectively communicated. Authority can manifest itself only through effective communication.

---

[2] In actual fact the process of communication is somewhat more complicated than we have just indicated. We have telescoped at least two aspects of it into those that we have made explicit. Consider this schema of the sequence:

Idea → signals sent → signals received → interpretation

The signals sent and signals received may be sounds or written words, etc. They are purely technical matters and we ignore this part of the process. We assume that the signals received are the same as those sent. Of course "noise" may enter the sequence and interfere with the signal flow, but we look upon that as an engineering problem and ignore it for our purposes.

AUTHORITY, COMMUNICATION, AND ORGANIZATIONAL ACTIVITY

Authority and communication are the life blood of the organization. Through them orders are given, decisions communicated, and actions taken. One might say that it is through authority and communication that a mass of human beings is made into an organizational machine.[3] The analogy of a machine would be especially apt where the authority and communication relations are functions —that is, where in all cases there are one-to-one correspondences between the superior's alternatives and the subordinate's alternatives, and between the sender's messages and the receiver's interpretations, and where in all cases the perceptions of the relations are identical for all parties involved.

Let us look at the machine analogy again for a moment. Suppose someone is trying to make a machine in which the parts are human beings. The reason he may want to do this is that human beings operating as a coördinated group may be able to accomplish more than when they work in isolation. A good machine is a set of precisely coördinated parts. When some parts are worked, other parts move and operate as a consequence. For example, in some machines, if we push certain buttons, certain wheels will be expected to turn. In the same way we can create machines using humans as parts. By imposing or creating authority and communication relationships we obtain a mechanism whereby when one in authority makes certain choices (pushes certain buttons) and communicates them, certain activities follow. If these relations are one-to-one correspondences, the activities that follow will be exactly those that are expected by the authority (the button pusher). This is, of course, similar to what we expect in a perfectly functioning inanimate machine.

However, it would be usually quite foolish to construct such an organizational machine. Given the properties of the parts involved we might construct much more effective machines—machines in which the parts utilize their capacity to store and sift information, to make choices, and often to make sensitive and complicated judgments. We shall see as we proceed that if we utilize these capacities of at least some of the parts of our potential organizational machine we can, in most cases, construct much more effective machines than otherwise.

---

[3] Keeping in mind the qualifications in the last chapter about the machine analogy.

The main reason the simple mechanical machines we have described do not lead to the most effective organizational forms is that we have not taken into account such unique capacities of individuals as their capacities to store information, to gather information, to assess situations, and to make recommendations based on such assessments.

Also, the complete dictatorially directed type of organization circumscribes considerably the expansibility of the organization. A dictator who gives out detailed orders with respect to every activity that is to take place can operate an organization of only a very limited size. Beyond a certain point the mere task of detailed decision-making and detailed order communication becomes too much for any single individual. Obviously for organizations beyond some rather small size authority must be shared. One of the main problems of building an organization is to determine how, exactly, it should be shared.

What we have in mind can perhaps be seen more clearly if we consider both simple and complicated forms of the decision-activity nexus.

The simple form can be summarized with the aid of a four-quadrant diagram. This is illustrated in Figure 23.

Quadrant A shows the relation between S's alternatives and the messages that S could use to communicate his alternatives. We shall assume in all cases that these relations are simple functions—that is, one-to-one correspondences. Thus for each alternative that S may choose there is one message that he would use (shown on the abscissa) to communicate his choice to I (the subordinate). The line *CM* shows all combinations of alternative choices and their appropriately associated messages.

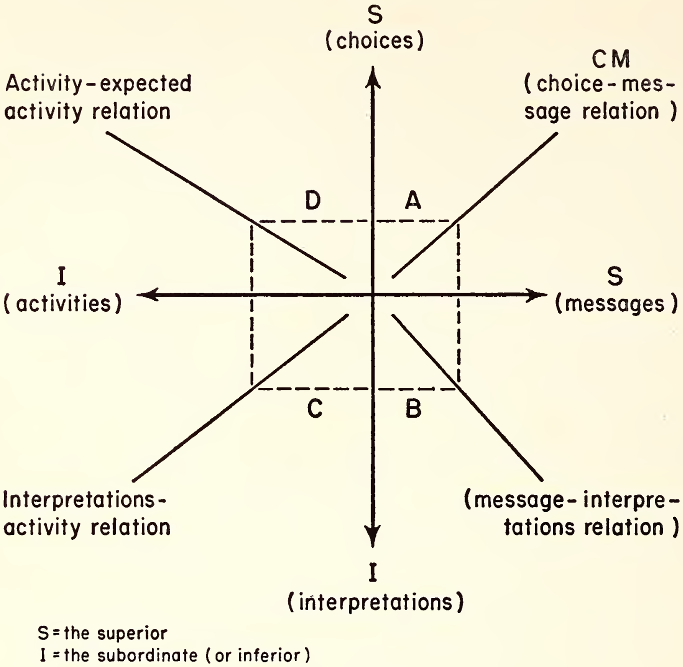

S
(choices)
CM
(choice-message relation)
Activity-expected activity relation
D
A
I
(activities)
S
(messages)
C
B
Interpretations-activity relation
(message-interpretations relation)
I
(interpretations)
S = the superior
I = the subordinate (or inferior)

Figure 23.

In quadrant B we have the relation between the messages sent by S and the interpretation given to them by I. In quadrant C we show the relation between the interpretations of I and I's actions. And quadrant D closes the picture by showing the relation between the choices of S and S's anticipated actions by I as a consequence of these choices. If the actual action by I is the same as the anticipated one, then we have an equilibrium situation requiring no adjustments. Otherwise, we might expect S to try to adjust the situation by making additional choices and communicating additional orders.

As we have already indicated, there are limits to a dictator's capacity to make and communicate detailed decisions with respect to every activity that is to be performed. Beyond some scale of operations the number of decisions becomes too large and the dictator must find some means of reducing his burden. Once he does so, the simple chain of choice, communication, activity, and the matching of activity with initial choice is broken.

The dictator or director can, however, make his choices and decisions on a higher level of abstraction, given the appropriate organization arrangements. Instead of choosing a specific alternative he may choose an alternative segment (i.e., a portion or subset of the alternatives in his field of action) and leave it to someone on a lower level of authority to use more detailed information and in turn choose either a smaller alternative segment or the specific alternative. In other words, if one is faced with choosing among too many alternatives he can reduce his task by dividing up the alternative range into suitably few segments. Then, by utilizing more general information, he can reduce his task to the choice of a segment. Of course, he must proceed to communicate his choice to a subordinate, who in turn must make the ultimate specific choice, or whose subordinate in turn must do so. In any event, by breaking up the alternative range into segments, and simultaneously increasing the layers of authority, it is possible to increase the scale of operations that can be handled by those in authority.

An alternative approach would be for the top authority not to choose the segment himself, but to communicate a rule or policy that would enable his subordinate to choose the segment. In all cases each layer of authority must communicate some procedure, although a degree of discretion may be permitted within that procedure, for the selection of the segment, the subsegment, and finally the detailed alternatives which will lead to performance. In sum, each subordinate would use the rule of selection plus the additional local information of a more detailed nature to make the choice at his level and to pass it on to the next echelon of decision-makers.

Through the use of such procedures increased efficiency could be obtained by selecting for the various roles players who have the specialized capacities and could utilize the detailed knowledge to make the most effective decision at that level. For example, one of the choices to be made may be the determination of the means of transport of some raw materials. It is clearly not necessary for the president to know all the alternative rates for all the possible means of transport. This matter can be handled more effectively through the employment of a professional traffic manager who has the requisite specialized knowledge and capacities and who is provided with a policy or rule of selection by higher management.

Let us now look again at the decision-activity-control nexus from the viewpoint of nondictatorial (in the detailed sense) decision-making. If the top authority chooses only a segment of the alternative range, or only a rule for determining that segment, then he can no longer match his choice with the activity that is finally performed at the end of the chain. There are two reasons for this: First, there can no longer be a one-to-one correspondence between his choice and the activity—a great many activities may be consistent with the large alternative segments. Second, even if he wanted to match this consistency he could not do so without examining the detailed information employed at every echelon which finally resulted in narrowing down the alternative range to a specific decision. Clearly, the top authority could never take the time to do this with respect to every detailed decision that is made in the organization. He could, however, introduce a control system of spot-checking such detailed decisions with initial policies, in order to see that in general the organizational choice procedures are of such a kind that detailed choices made and activities carried out are in fact consistent with top policy pronouncements. We can also see that in general the top executive of a large organization must lose control of detailed activities. Schemes may be worked out so that the organization as such can be said to be in control of the situation, but no single individual is directly in this position.

We will have more to say on the distribution of authority problems in a later chapter.

CONCLUSIONS

We have attempted to sketch the elements of the authority-communication mechanism. Through authority and communication the activities of different individuals are linked, and the capacities of some individuals (those who have authority) are extended. Also, it is in part through the authority-communication mechanism that a group of separate individuals can become an organization.

In a small organization it may be possible for the top authority to check in detail on the outcomes of all of his decisions and orders, but in a large organization it is impossible for a single authority (say, a dictator) to give all the orders, and to maintain the detailed connections between his decisions, orders, and communications and the subordinates' interpretations and actions, in such a way as to compare the latter with his (the dictator's) anticipations. The larger the organization, the less the amount of detailed control that the top authority as an individual can retain over activities. Therefore, in organizations larger than some very small size, authority must be shared. Others in the organization must be given some degree of discretion. Performance in detail cannot be strictly determined by decisions at the top. The area of discretion is usually reduced in part by the introduction of rules of choice and standardized procedures. But such techniques do not entirely remove the possibility of decisions and activities by subordinates that are unsatisfactory from the superior's standpoint. On the one hand this creates the need for control, a matter that we have touched on previously. On the other hand, it helps to explain the concern with the motivations and incentives of subordinates, and with the connection between motivation and performance. We turn our attention to these matters in the next chapter.

12

Decision-Making and Motivation

CHOICE SETS, ACTIVITIES, AND CONSEQUENCES

In the previous chapters we have considered the basic characteristics of an organization. We now proceed to the motivational elements—the binding stuff of the organization.

Motivation is intimately connected with the agreement to participate within an organization, with the day-to-day decisions and operations, and especially with the participants' perception of the relations between decisions, activities, and their consequences. To put motivation within its proper context we consider first the connection between choices, activities, and consequences.

Let us recapitulate briefly what we have thus far. The firm as an organization is to be looked upon as a collection of related roles. Within these roles there are individuals who spend a portion of their time as role-players. They interpret their roles in a certain way, keeping in mind that in many cases more than one interpretation is possible and feasible. An interpretation of a role is made up of a collection of routine and contingent fields of action. These in turn are to be looked upon as sets of exclusive alternatives. Each alternative, if chosen, implies an activity of some sort. When a role-filler plays on a given field of action he does two things simultaneously: he chooses an alternative and he acts in accordance with his choice. Thus, a *move* by a player implies the choice of an alternative, an act in accordance with that choice, and the consequences of that act.

Consider a single *play* of a role. By a play we mean an operation on all the fields of action within a role interpretation in accordance with the relative frequencies of the fields of action in that interpretation. The result of a play is a choice set, by which we have in mind all of the decisions made out of the various fields of action within the role interpretation. In sum, we visualize that the outcome of a play is a choice set, a set of activities, and finally a set of consequences flowing from these activities. The elements involved in the play of a role are illustrated in Figure 24. The fields of action constituting the role interpretation are A_1, A_2, and A_3. By the prefix 2 in $2A_3$ we mean that A_3 will be moved on twice in the same period that A_1 and A_2 are moved on once. Therefore, from A_3 flow two choices, two activities, and two sets of consequences.

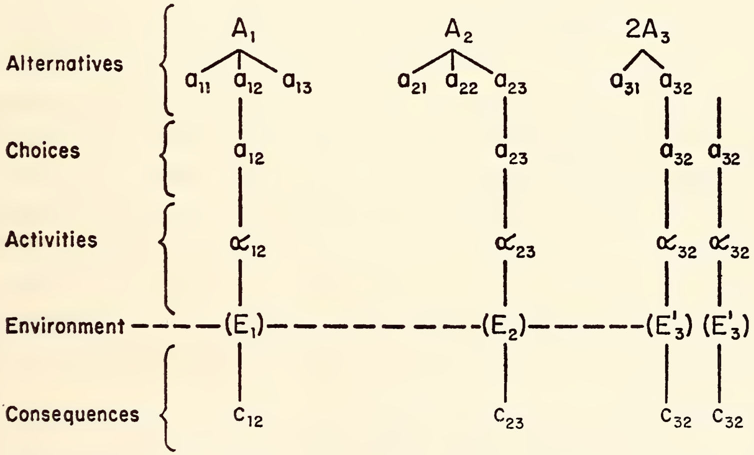


Figure 24. Diagram of a play of a role.

The relation between activities and consequences is not a fully determined one. That is to say, we must not assume that the same activity at two points in time will necessarily lead to the same consequences. It will depend on the environment. Therefore, the consequences may be said to depend on two elements: the nature of the activity and the nature of the environment.

The diagram indicates what is involved in a play of a role. But it does not indicate how the decisions are made in the first place. In other words, there is nothing in our discussion to suggest any propositions about behavior. Of course, if we did put some behavior propositions into our scheme, we could possibly have a completely determined system. Let us see what we might have to know in order to determine the outcome. In the first place we would have to have some behavior propositions in connection with the choice of a role interpretation. Given the role interpretations perceived by, and open to, a player, what specific role interpretation will he choose, and what will determine his decision? Will he choose the one that will maximize his leisure, or the one that will minimize his responsibility, or the one that will lead to a more desired role, or what? Quite clearly there are a large number of behavior assumptions that we could make, and for different sets of behavior assumptions we might get different results. The main point to keep in mind in connection with the following sections is that behavior is determined, at least in part, by the motivational elements in the picture—by the interaction of the values and goals of the participants and the perceived inducements offered by the organization that enable the participants to pursue their goals in accordance with their values.

Similar things may be said with respect to the play of a given role interpretation. The player may play within each field of action so as to minimize responsibility, or maximize the profits of the firm, or maximize his recognition by his superiors, and so on. The choice of a role interpretation may be looked upon as a long-run strategy, whereas the specific playing of a role is in the nature of a short-run strategy. In any event, it should be clear that (1) a number of behavior assumptions can be chosen, (2) for each collection of behavior assumptions we would have a different model of the functioning of an organization, and (3) the behavior is, in part, determined by the motivational scheme in force.

For present purposes we need not designate any specific rules of selection or criteria for choice that are presumed to be used by the role-players. In fact each player will have to adopt, explicitly or implicitly, some decision rule (or set of rules) in order to choose (1) his role interpretation and (2) a specific alternative out of each field of action. The rules or criteria may be formal and specific, or they may be general and perhaps vague and themselves subject to a number of more specific interpretations. Broadly speaking, interaction between roles, and therefore role-players, will normally determine whether the role interpretation and the decision criteria employed by a given player are organizationally permissible. Indeed, an adequate organizational control system should lead to those role interactions that inform players whether their behavior is in accordance with presumed organizational objectives and constraints. In general, therefore, there should be some degree of consistency between the tacit understanding of organizational objectives (or the explicit pronouncements of organizational policy by the top authorities) and the interpreted rules of selection at various levels and points in the role structure. For any specific organization there is the problem of determining whether in fact such consistency exists. But this problem we leave to a later stage of our discussion.

It is of interest to note that not every move within a field of action need lead to positive consequences, especially where the fields of action involve coöperative or competitive relations with other role-takers. Consider the effect of a man voting in a committee. If he votes on the losing side, there may be no consequences whatsoever to the fact that he made a choice within that particular field of action. Its consequence is zero. Similarly, if an individual competes with another individual within the organization setup but loses in the competition, again the consequences may be zero. Of course, such votes or losses in competitions may have long-run indirect effects. All that we are trying to assert now is that we must recognize the possibility of zero consequences.

The consequences of some decisions and acts will not be zero. The sum of all the consequences of the activities of the role-takers will determine the operations, and the success or failure, of the firm. We may divide the consequences into two categories: those that are external to the firm and those that are internal. The promotion of a man from one position to another, a change in the duties of an individual—these are examples of consequences entirely within the firm. Consequences based on the acts of production and delivery of goods to outsiders have more complicated effects. First, they change the amount of goods both inside the firm and in the outside world. Second, and simultaneously, they increase the income and profits (or losses) of the firm. Clearly, for the individual firm it is in some sense the aggregation of the consequences of the choice sets that determines the profit, losses, contraction, or expansion of the firm during the period.

There is another side to the consequences of the choice sets: the consequences as seen by the individual participants from their *personal* viewpoints. To them some of the consequences may imply the possibility of promotion, a higher share in the payoff, and so on. In other words, some of the consequences are from their viewpoint closely connected with their objectives and motivations.

MOTIVATION AND INDIVIDUAL DESIRES

By motivation we have in mind the relationship between the privileges and disutilities (i.e., rewards, penalties, etc.) that are associated with a role, or with particular moves in playing a role and the desires of role-players or prospective players. If the rights, privileges, etc. (we may call the collection of such attributes the *motivators*) associated with a move help to fulfill a player's desires better than those associated with any other move, then he is presumed to be motivated to make that move.

Motivation has meaning and significance because conscious human behavior is goal oriented in large part. Individuals have desires of various kinds and perceive various entities, activities, or associations as means of helping to satisfy them. In many cases these desires are sufficiently strong so that the individual is willing to give up something, or to act in a particular manner, in order to obtain some of the means of want satisfaction. This, of course, is the economic basis for exchange. If the means for want satisfaction are marketable entities, such as food and shelter of most kinds, then money rewards, or rights that could be translated into money, will serve as an inducement to performance. But some of the means of want fulfillment are not marketable entities and hence nonmonetary inducements must also be taken into account. Indeed, the contribution of the industrial psychologists such as Elton Mayo and others has been to emphasize the significance of the nonmonetary and nonmarketable inducements in determining performance.

Even in the case in which monetary inducements could achieve the desired result it may sometimes be to the interest of the organization offering the inducement to use nonmonetary inducements. For example, some of the nonmonetary inducements may turn out to be less expensive. Also, beyond some point a participant in an organization may become more interested in nonmonetary inducements—status, career, prestige, peer approval—than in additional monetary inducements. In other words, as an individual receives more and more money for his efforts, the wants that could be fulfilled through marketable means become satisfied to such an extent that those wants that could not be so fulfilled begin to appear the more urgent.[1]

---

[1] Some of the studies and surveys made by industrial psychologists reveal that when employees rank the important attributes of a job, salary is usually toward the middle of the list. Job security, possibilities for promotion, and the nature of work usually come higher in such lists. It is to be noted that these studies force workers to rank things in order of importance to them whether or not they can, in fact, make such distinctions. Also, the results are biased, in part, by the fact that in a mobile labor force workers have already chosen the jobs in which the salary appears to be satisfactory to them. It is the other elements in the situation, especially those that cannot easily be determined prior to participating in an organization, that are likely to be unfulfilled. Since at any one time not all of the desires of participants are going to be fulfilled to the same extent at the margin, the fulfillment of some desires will appear more urgent to some of the participants than the fulfillment of others. In other words, we can apply the usual marginal analysis to the motivation problem just as we apply it to other situations in which there are competing desires to be fulfilled. See, among others, Milton L. Blum, *Industrial Psychology and Its*

In what follows we shall examine decision-making both from the viewpoint of the player and from that of the organization. In discussing motivation from the player's viewpoint it may be helpful if we isolate and keep in mind these three elements of the situation: (1) the player's desires, values, and goals, (2) the player's perception of the alternatives available, i.e., his perception of possible moves, and (3) the player's perception of the connection between these moves and rights, rewards, and penalties associated with the consequences of the moves.

THE MOTIVATION TO PARTICIPATE

There are at least three general choice areas in which motivation will play a part: (1) the acceptance of the role, and the shift from one role to another, (2) the choice of a role interpretation, and (3) the choice of alternatives within a field of action. The first choice area involves the decision to participate within a specific niche in the organization. The second involves the way one chooses to participate within that niche, and the third involves the specific moves within a role interpretation. Clearly, in all three areas elements of choice enter, and in all three incentives and inducements will play their part.

It is often useful to separate (1) from (2) and (3), i.e., to separate the motivation to participate from the motivation to perform. In the rest of this chapter we will emphasize the motivation to perform, in part because it is glossed over in economic theory. But to complete the picture we consider briefly the motivation to participate.

In the previous section we looked upon motivation in terms of an exchange transaction. On the one hand we visualized an inducement (reward, or right of some sort) as being attached to a role, and on the other hand we pictured this inducement as a means of want fulfillment for a player, who in accepting the inducement was

---

Social Foundations, Harper & Brothers, New York, 1956, pp. 59 ff., on the alleged overemphasis of monetary rewards. See also J. A. C. Brown, *The Social Psychology of Industry*, Pelican Books, Baltimore, 1954, chap. 7.

willing to give up something in return. We may easily analyze the motivation to participate in these terms.

To begin with, we recall that an organization is a role-creating agency. In order to fill the roles created, some rights and privileges have to be associated with them so as to attract potential role-players. For example, with most roles there may be associated a salary, some specific status, career opportunities, and so on. It is important to keep in mind that not all roles in organizations have salaries attached to them and yet there is often no scarcity of potential role-fillers for them. For example, in some cases boards of directors of firms or trustees and even officers of noncommercial organizations do not receive salaries but the other want-fulfilling means that they provide such as that of status, prestige, and peer approval will often induce people to accept these roles.

While on the one side certain "motivators" are required in order to induce people to fill roles, on the other side the individuals that fill them are usually expected to make a contribution to organizational activities in some way or other. Whether or not a person is willing to fill a role depends on whether he gets out of it more than he has to give up. The usual economic explanation is in terms of utilities. A person will accept a role if the expected utility of participating is greater than the utility forgone by not doing so. The utility forgone will depend on the alternative opportunities that the individual perceives to be available to him. Thus, the other roles that he believes to be open to him and the value of the inducements associated with those roles will determine for him the utility forgone. In sum, we can analyze the motivation to participate in terms of the usual analysis of exchange.

At this juncture we may recall that the inducements associated with various roles are, in part, created by role-playing within the organization. The consequences of the choice sets and activities of the players are an element in determining the earnings of the firm and therefore setting limits to the possible inducements that the organization could associate with various roles. In other words, the contributions of the players, determined by the way that they play their roles, determine the aggregate payoff of the organization and hence the pie out of which the inducements must be cut, so to speak. We have here a condition for the survival of an organization. The sum of the contributions of the role-fillers must be at least large enough to enable the organization to associate with various roles a set of inducements that will entice competent role-fillers to participate in the organization.[2] Thus we must consider the motivational scheme the binding stuff of the organization. For it is this scheme that entices into the organization people who are capable of contributing sufficiently to its activities so that inducements to attract their continued participation may be offered.

MOTIVATION IN ECONOMIC THEORY

A good deal of the literature on motivation is concerned with motivation on the job, and to some extent with nonmarketable inducements and noncontractual motivational aspects. To this we may contrast the very limited part that motivations play in traditional economic theory. In the theory of the firm it is assumed that the owners seek to maximize profit and that the owners of factors of production seek to maximize their income. While conditions of work may be of importance to an employee they are assumed to be known at the outset to the worker and to remain fixed during the contractual period of employment. Furthermore, it is assumed that the nature of the tasks to be performed are fully understood by employer and employee and therefore motivations as such play only a limited role. If the conditions of work and the nature of the job to be performed are given, then there is not much room in such a theory for an intrafirm motivational system since the theory assumes that the jobs to be performed will in fact be performed in accordance with the prearranged employment contract.

In other words, microeconomic theory is concerned only with the inducement to participate but not to perform. Performance in accordance with a prearranged contract is taken for granted. The entire matter is viewed in terms of an exchange transaction on the basis of full information. The seller of labor presumably knows what he gets in terms of salary and other rewards. At the same time he is presumed to know what he has to do in return for them. The transactions are between the firm and sellers in the market place but there are no intrafirm transactions. Thus, within the conventional framework of economic theory very little can be said about the motivation to perform within the firm. A similar situation exists with respect to leaving the firm. The worker is presumed to know the alternative work opportunities elsewhere, and if one of them is superior to his current opportunities he will supposedly leave. Once again, he is presumed to leave one mode of employment for another on the basis of full knowledge of the inducements and contributions involved.

---

[2] Cf. James G. March and Herbert A. Simon, *Organizations,* John Wiley and Sons, New York, 1958, pp. 84 ff.

THE MOTIVATION TO PERFORM, KNOWLEDGE, AND DISCRETION

The agreement to participate in an organization is rarely complete and precise on both sides. For example, when someone joins a firm there is usually agreement with respect to his salary, the hours of his work, and the nature of some of his main duties. The intensity with which he is expected to carry out his duties will probably not be part of the formal agreement. Similarly, other elements in the work situation such as some of the specific details of his duties and the opportunities for movement within the organization may not be specified in advance. Many of the benefits and inconveniences that a participant will experience in the organization may be incompletely revealed at the outset. Usually the initial period of participation within the organization is a period of exploration and learning with respect to the nonformally contracted for benefits, rewards, expected performance, and inconveniences. The role-taker is able to perceive new areas of potential behavior and connections between these and the rewards and penalties that may be associated with them from his own point of view.

In addition, the nature of the work is usually such that some degree of discretion has to be left to the employee. As a consequence there is, by the very nature of the situation, no way for him to perform precisely in accordance with his contract since the contract does not specify the performance in exact terms. In most cases the organizational setup must be such that some variation in performance is permitted to many members in the organization. We have discussed in previous chapters various reasons why it would be impossible to establish a completely mechanical and rigid method of carrying out organizational activities. If all orders given could be responded to in a precise way without involving any interpretation or discretion, then, of course, motivation would not be significant. Similarly, if all information necessary to make decisions were always available to the top authorities, precise orders could be given and once more there would be no question of motivation. Again, if all the rules of selection given to subordinates required no interpretation in their application, motivation would not enter into consideration. Since, in fact, given the nature of human beings, all of this is not the case, it follows that incentives and penalties will have a place in determining individual decisions. How a participant performs will depend not on the prearranged contractual conditions but to a great extent on the motivational system that he faces in the process of performing. We have here an element that is not contained in the economists' theory of the firm.

In this connection consider briefly some of the incentives that organizational participants presumably respond to other than salary. Employees are often said to be concerned with such matters as (1) job security, (2) promotion and career possibilities, (3) the nature of the work and the satisfaction to be obtained from the work itself, (4) the status associated with the job or profession, (5) peer approval within the organization and the approval of professional peers outside the organization, (6) the approval of superiors within the organization, and (7) since most jobs provide opportunities for social relationships, the social satisfactions to be obtained from the role. This list is, of course, by no means exhaustive. It is simply a sample of wants that are sometimes fulfilled in the work situation —wants often mentioned by industrial psychologists. The role structure of the organization and the contractual arrangements made at the outset will determine the extent to which these various desires can be fulfilled by participants in the organization.

We have indicated that the motivational system is to be viewed as being operative at the same time that the role-taker is performing within a role. This implies that the participant is able to connect his activities with the possibility of want satisfaction on the job. In other words, the role-player makes certain choices; in doing so he is able to perceive the connections between the alternatives open to him and the satisfactions provided by the organization and associated with each of the alternatives.

In selecting between alternatives the participant will follow certain rules or other guides established by the organization, but there will usually be an area of some discretion. That is, the organizational rules will not determine the precise alternative that is to be picked in all instances. The participant will choose between some alternatives for which no exact rule exists, or between fields of action in which he feels free to disregard some of the existing rules.

The participant's perception of the situation now enters in two ways. First, his perception of his role and of organizational expectations connected with his role will determine in part his area of discretion. Second, his perception will determine the connections that he sees between different alternatives, their consequences, and how much they help satisfy his desires as translated by his value scheme. Thus, we have here one side of the motivational picture, the alternatives-consequences-want satisfaction chain for the participant.

The amount of discretion permitted will depend on the controls within the organization and on the extent to which participants are presumed to be loyal and able to make choices in accordance with organizational policies and objectives. The greater the degree to which people are presumed to be able to make choices in accordance with organizational objective and policy, and the greater their desire to do so, the less the control procedures that are necessary. Since controls, especially of a very detailed kind, are likely to be costly, it follows that means of inducing loyalty to organizational objectives, and behavior in accordance with such objectives, should be desired by the organization. Clearly behavior in accordance with broad organizational policies rather than within detailed rules involves a much higher degree of discretion. We shall see later that at least some members of the organization will have to be given such discretion. Hence the motivational system is significant with respect to performance because it will determine how the discretion is employed.

To summarize, we have seen that one attribute of a role is to have rights and privileges associated with it, and that the nature of these rights and privileges is usually such that they serve as an inducement to participate and to perform. Now, not all of the rights and privileges are contracted for at the outset or always made explicit prior to a player's entry into the organization. Some of them, especially the more subtle elements, will be discovered while participating within the organization. Some, too, will depend on performance or on other events that will take place while an individual is a member of the organization. The very right to continue in the role will usually depend on good performance, and the rights to increases in salary, promotion, etc., may also hinge on some index of performance or service.

Given the fact that some rights and rewards will be connected with performance, we can now delineate other basic elements of the motivational scheme. First, we have the connection between rights, inducements, and the wants of the individual. The rights associated with a role or the rights that become manifest upon certain moves will serve as an inducement toward certain types of behavior because they will enable the player to satisfy some of his wants, directly or indirectly. Second, there is the connection between inducements and performance. The nature of the rights, and the inducements that they become, will determine the nature of performance and its desirability as viewed by the organization. Third, the inducement-performance connection depends on the individual's perception of that connection. The extent to which a player perceives that he could achieve additional rights, or maintain existing rights, by a higher level of performance will determine the extent to which he is motivated to perform in that manner. Fourth, from the organization's viewpoint we note that the use of different types of inducements is open to the organization, but the effectiveness and costs of different inducements will depend on player selection and organizational structure.

MOTIVATION AND ROLE STRUCTURE

The motivation scheme will operate through the role relations. Since a great many motivational elements depend on decisions made by one person with respect to the performance or some other participational element of another person, the system of role relationships will clearly determine when, by whom, and how such assessments are made. For example, player selection is determined by those in the player-determining roles, and the nature of the role structure will determine the way in which player selection is perceived by a participant. The same will be true with respect to the payoff-determining roles. It will make a difference whether these roles are close to the participant or far away, whether there is frequent or infrequent contact between individuals in such related roles, whether there is a long decision-making chain remote from the individual that determines his payoff, advancement, conditions of work, and so on, or whether these decisions are made with few intervening steps in his immediate environment. The motivational effects of decisions made at some home office far away from the scene of performance usually have less impact than those of decisions made close to the scene of performance. In addition, the opportunities for advancement will clearly depend on the nature of the role structure. A role structure with a long hierarchical chain and one with very few hierarchical steps obviously present different advancement opportunities. The rate of change of players in and out of roles, which is connected with the nature of the role structure, will also influence the extent of opportunities for advancement.

Suppose that promotion to a better post is determined by proficiency in the current job. This criterion will operate as an incentive for a participant to do well in his job if he is interested in his promotion—and if he perceives that the connection exists. In this case, the participant's role is connected with the player-determining and role-determining roles in the organization. Someone in a player-determining role will have to judge whether this individual is doing well on his job, and whether he should be the one to be chosen for a better post. If this connection is not perceived by the participant, then, of course, it can have no influence on his output and performance even if, in fact, the organization promotes on the basis of proficiency in the current job. On the other hand, if promotion is on the basis of seniority, family relations, or internal politics, and is understood to be so by the participant, his performance will not be influenced by the prospect of promotion. Indeed, if promotion is based on internal politics, performance, from the viewpoint of the organizational objective, may actually be inhibited because it induces people to devote their energies to internal politics rather than to the improvement of performance. A similar analysis may be made with respect to some of the other motivating elements.

Consider one more case, that of peer approval. It has been reported in a study by Melman[3] that in the Standard Automobile Company in Coventry, England, workers are paid in accordance with group performance rather than individual performance. As a consequence, any event that is likely to decrease the group performance is a disturbing factor and any member of the group who inhibits group performance, of course, meets with group disapproval. Here is a case in which the desire for peer approval operates in the direction of increasing performance from the point of view of organizational objectives. The general role relations are not too difficult to trace. On the one hand, we have the coöperative relationships between members of the same group; on the other, the relation between these and individuals in the payoff roles of the organization.

Other studies have shown that production rates often vary with those of neighboring workers. This may have something to do with the desire for peer approval, but in addition the output of neighboring workers provides clues about the nature of acceptable work standards where these are not provided by other means.[4]

---

[3] Seymour Melman, *Decision Making and Productivity*, John Wiley and Sons, New York, 1958.

A complicated variation of the motivational aspect of the desire for peer approval enters the scene with the rise of professionalism and the employment of professionals in organizations. The professional is often much more concerned with the approval of his professional peers than with that of his organizational peers. Therefore, a significant inducement considered by a professional in the decisions to participate and to perform is whether the organization provides an environment in which he can do the type of work that will receive the approval of his professional peers.

There are many facets to this situation, some of which are related to the role structure of the firm and the role relations within it. At the outset we note that the professional plays two roles simultaneously—one as a member of a profession and the other as a member of an organization. If possible, professionals prefer to work in the environment of their fellow professionals. Hence, if the firm is to provide an environment in which the professional can obtain contact with and possible approval of his professional peers, it may have to be set up in such a way that members of the same profession are in the same departments, or in close contiguity with each other. This consideration can be a constraining element in determining the role structure of the organization. There is also the possibility that professional standards of behavior may conflict with organizational standards. For instance, the standards of the architect in solving a certain problem may be quite different from those of the finance officer, who may be concerned mostly with the cost aspects of the situation. On the other hand, if professional standards of behavior are consistent with organizational requirements, then part of the control problem is solved for the organization by the profession. In any event it is clear that members of an organization who are in a position to assess the work of their professional colleagues must take cognizance of professional values and standards. In general, the effectiveness of the motivational scheme in those cases where professionals participate in an organization will depend on the role structure in which the professional roles are embedded, and on the role relations between the professional roles and others in the organization.

---

[4] March and Simon, *op. cit.*, p. 56.

The potential output-reducing effect of the desire for peer approval can be observed where workers feel that if one person performs much better than some norm, his performance will become the new standard for the employer and will result in a "work speed-up." This is the well-known phenomenon of the fear of "rate busting." Here the desire for peer approval will induce the participants to limit rather than to extend their performance.

Perceptions are evidently crucial in this area. On the one hand, we have the participant's perception of his desires and how they are to be fulfilled; on the other hand, his perception of how the consequences of his choices can help to fulfill his desires. At this point the motivational scheme enters the choice-performers mechanism, because the perceived consequences of his choices will depend on the organizational responses to his behavior. For example, if his choice is made in such a way as to impress his superior, who could further his desire for promotion, then clearly it is the participant's perception of the superior's capacity to respond that motivates him to choose as he does. Hence, if the organization is to influence the motivation of an individual, it must influence that individual's perception of the connection between alternatives, consequences, and the fulfillment of the participant's desires.

That motivation depends on player selection can be seen readily. The same set of rights associated with a role will mean different things to different players having different wants. By the same token, the same rights will induce different players to behave differently. Thus player selection determines which players are chosen, the want characteristics of different players, and hence the extent to which the players actually chosen will be moved by the rights associated with the roles for which they are chosen. If an organization cannot manipulate the rights associated with various roles, say, in the short run, it can at least be sufficiently careful about player selection so that those selected are the ones who perform best given the inducements determined by the role's rights and privileges.

SHORT-RUN VERSUS LONG-RUN MOTIVATORS

For some analytical purposes it may be useful to distinguish between short-run and long-run inducements. By short-run inducements we have in mind those connected only with the activities of a given unit period. By long-run inducements we have in mind those connected with the activities of more than one period and not involving the simple summation of the inducements for single periods. For example, a monthly salary would be a short-run inducement since it is connected only with the activities during the month in which they took place. A promotion scheme, on the other hand, will usually be a long-run inducement if the reason for promotion is connected with performance over a number of periods.

Important short-run inducements to be considered are the following: salary, peer approval, work satisfaction, and prestige of the organization. The basic aspect of short-run motivation is that the rewards do not depend on continued performance. Of course, the form of the inducement does not in itself completely determine whether it is a short-run or long-run inducement. For example, a system of automatic salary increases which depend on length of service is a long-run inducement although it is salary that is involved. Some of the long-run inducements that might be considered are the promotion scheme, bonuses, profit-sharing arrangements, career opportunities within the organization, seniority rights, retirement benefits, and so on. Most of the titles used suggest the nature of the inducement and the motivation involved; hence we need not go into detail.

A major question is why an organization should offer long-run inducements. Several important reasons come to mind.

In many roles there is a breaking-in period and often a training period, during which performance is either negligible or far from being at its peak. Thus the firm may be said to be making an investment in an individual in order to establish him effectively in his role. A return on this investment can be achieved only if the person continues as a role-player for some time. Since people are free to quit their jobs at will, or at least on very short notice, it is necessary for the organization to have long-run inducements associated with the role in order to obtain from the player a reasonable period of continued participation. The longer the time period of participation necessary for the firm to get a sufficient return on its investment, the greater the inducement necessary. Clearly, different types of inducement will have different degrees of effectiveness in achieving their objectives. Which type of inducement is most effective will depend on the preference scheme of the players chosen, the alternative opportunities that they have available to them at various junctures in their career, and the general economic environment. In situations or areas where the supply of labor is much greater than the demand, that is, where jobs of a certain type are scarce, very limited long-run inducements, or even none whatsoever, may be sufficient. On the other hand, if jobs are plentiful, as the player becomes more valuable to the firm, and perhaps at the same time to other firms, his rewards for performance must increase to such an extent that it does not pay for him to shift to another firm. This will usually mean that his income or other rights, especially rights in future rewards, will increase as time goes on, and the loss to the firm upon his withdrawal will also rise as he becomes more valuable to it. In some cases a promotion scheme or career pattern will do the job; in others a system of increased participation in year-end bonuses or profit-sharing or in retirement rights, and so on, may be more effective. These are matters of detail that cannot be determined on an a priori level.

Often it is desirable for the inducement to be of such a kind that it creates a close sense of identity between the achievements of the firm and the achievements of the role-player as a member of the firm. In other words, it is desirable for the role-player to feel that any of his moves that help achieve the firm's objectives are moves in his own interest. Many long-run inducements can be of this kind—for example, a profit-sharing scheme. If the role-player behaves in such a way as to increase the company profits, he will simultaneously increase his own income. Similarly, a promotion procedure that depends on a man's contribution to the profits of the firm or to some activity directly or indirectly enhancing these profits is also of this type. If the firm's objective is not profit maximization but growth within a profit constraint, then a promotion scheme that depends on the growth of the firm will also serve as a motivational scheme that increases the identity of interests between the role-player and the organization. Indeed, often the faster the growth of an organization, the stronger the sense of identification engendered in participants.[5] This last may be especially true for the somewhat higher echelons of the work force, who are likely to have a lively sense of anticipation of the sort of opportunities that will be generated by growth. It is often found that executives are paid in accordance with their span of responsibility—roughly the number of people accountable to them directly or indirectly—rather than with the profitability of the operations with which they are connected.[6] The greater the rate of growth, then, the greater the size of the organization and the greater the chance that posts will be created in which the span of responsibility is larger.

In considering the types of inducements to use, the cost factor will usually be taken into account. In some cases long-run inducements may be less expensive than short-run inducements even though, apart from costs, the short-run inducements would do the job. For example, an apprenticeship system in which the apprentices do practically journeyman work but receive very low salaries in the early years may be of this nature. Other training programs may also be of this type.

Where the demand for labor is cyclical, it may be desirable for the firm to have long-run inducements in order to maintain the labor force during tight periods.

---

[5] *Ibid.*, chap. IV.

[6] See Norman S. Buchanan, "Theory and Practice in Dividend Distribution," *Quarterly Journal of Economics*, November, 1938, pp. 81 ff.

THE MOTIVATION PROBLEM FROM THE ORGANIZATIONAL VIEWPOINT

Since an organization has some choice in the rights that it attaches to various roles, it can select the types of inducements it wants to employ. Of course, the organization is not limited to a single inducement. Its problem is to determine the inducement mix, and the related organizational arrangements to make it most effective. The effectiveness of such arrangements will largely depend on (1) the opportunities for perception of inducements and deterrents that are available to participants, (2) the role structure and the allocation of discretion implied by it, and (3) the method of player selection. The method of player selection will determine the extent to which those picked to fill various roles are responsive to the inducement mix.

We have already indicated that a participant cannot be motivated by organizational arrangements that he does not know of or cannot perceive. Therefore, the organization must take cognizance of the extent to which participants are made aware of the system of rewards and penalties within the organization, the rules that determine advancement, the rules that determine the distribution of the payoff, the expectations on the part of the officers of the organization with respect to performance, standards, and so on. Not that it is always desirable from the organizational viewpoint for all information of this sort to be communicated to all participants. In some cases the availability of such information may be an inducement to anti-organizational behavior. For example, if promotion is on the basis of family connections but some of the participants are under the mistaken impression that it is on the basis of current performance, correcting this impression would not be in the interest of the organizational objectives, since this would eliminate one of the incentives for superior current performance.[7]

Now, in order to obtain a high degree of assurance that participants make choices that are desired from the point of view of the firm (or other type of organization) the firm can organize in one of two directions: It can attempt to narrow the area of discretion to such an extent that the participant is limited in his opportunities to choose undesirable alternatives; or it can permit a wide area of discretion but introduce incentives so that the participant chooses in accordance with organizational desires. The narrowing of the area of discretion could be achieved by giving precise and specific orders, by introducing rules of choice for the participant that are exceedingly detailed, or by establishing control procedures so that the individual's choices, when undesirable, would be easily corrected prior to the manifestation of the full consequences of these choices. Such techniques may be looked upon as attempts to avoid the motivation problem. For example, to the extent that a military organization depends on rigid discipline and obedience to detailed orders it is avoiding the necessity of achieving desired behavior from its soldiers by introducing incentives for them to behave correctly. All of this is not to suggest that attempts to avoid the introduction of incentives may not make sense in many cases. Indeed, it may be more efficient or less expensive to operate within a framework of limited discretion, detailed orders, and detailed controls than to attempt to introduce an incentive scheme to achieve the same results.

---

[7] Cf. March and Simon, *op. cit.*, p. 62.

Usually some members of the organization will make decisions on the basis of broad policies, while others will follow rules of choice and still others will operate under direct orders. The nature of the mix of these procedures will depend on the nature of the activities and on the organizational structure. What is clear is that those who make decisions on the basis of broad policy usually have a much higher degree of discretion than those who simply follow orders. The difficulty with discretion is that it may take some time to determine whether it has been used wisely and whether it has really been employed in the interest of the organization. As a consequence it is usually deemed desirable that those who are given discretion should identify themselves closely with organizational interests. Hence the frequent interest, on the part of those concerned with organization matters, with discovering motivational techniques that foster organizational identification.

The amount of discretion permitted in any one role is closely connected with that permitted in other roles. If very little discretion is permitted to most members of an organization, a great deal of it usually exists for those who have it. To make an extreme comparison, if everyone follows orders and only one person gives them, the person who gives them clearly has a great deal of discretion compared to the situation in which people coöperate voluntarily and without benefit of any authority relations. Thus, by changing the role structure and the authority relations in an organization we could alter the locus, distribution, and concentration of discretion. In other words, one of the choices open in some degree to an organization is whether to concentrate discretion mostly at the top or to spread it somewhat more evenly throughout the organization. Wherever a person could be held accountable for his actions, it is usually possible to develop rules for his behavior and to check performance against behavior within the rules. However, it is usually not possible to do so at all levels within an organization. Some people have to formulate rules in the light of organizational policies and objectives, just as others have to have a hand in formulating the organizational policies themselves. In those cases in which rules cannot be followed and accountability for behavior is exceedingly difficult, we would want to concentrate discretion in those roles where it can most easily be associated with inducements connected with organizational achievements.

We have already mentioned that some of the long-run inducements are effective in stimulating an identification of the role-player with the organization.[8] This is true of all inducements where there is a direct connection between organizational achievement and role-player achievement. Thus if promotion in some way depends on the profitability of the firm then this should help to establish an identity of interest between the role-player and the firm. A profit-sharing scheme is of a similar nature.

---

[8] In many situations leaving individuals some degree of discretion in their jobs may be desirable because it appears to increase the dignity and desirability of the work from the employee's viewpoint. In general people do not like to be very closely supervised in their activities. For example, in one study of office work it was found that supervisors of efficient units gave general rather than close supervision. Cf. D. Katz, N. Maccoby, and N. C. Morse, *Productivity, Supervision, and Morale in an Office Situation*, Survey Research Center, University of Michigan, Ann Arbor, 1950. See also Robert L. Kahn,

Now, what should an organization consider in order to have the best motivational scheme possible? Probably the most difficult part of the problem is that of assessment rather than that of determining the optimum scheme, given all the information. That is to say, with respect to a great many rights and potential rights that an organization could associate with a role, it would often be exceedingly difficult to determine how these rights would work out as inducements. In addition, it is often exceedingly difficult to determine the exact relation between inducements and productivity, especially where nonpecuniary rights and inducements are concerned. The extent to which piece rates influence production as compared with hourly rates may be calculable, but the degree to which one promotion scheme influences productivity in the long run as against another may be almost impossible to determine. A similar conclusion would apply to alternate systems of pension rights, stock participation arrangements, and so on.

The logic of choosing the optimal scheme is quite simple and will be sketched presently. However, it presumes that the assessment problem has been solved. We may look at this matter in two steps.

1. Given the organizational structure, we can associate various alternative sets of rights with the roles in question. Each set of rights will operate as a set of inducements which in turn will determine the output of the organization. Also, each set of rights involves a certain cost to the organization. Now, for a given role structure the best set of rights will depend on what is to be maximized. If net profits are to be maximized, the best set of rights is the one that yields the greatest difference between the net output produced and the cost of these rights for everyone other than the

---

"The Prediction of Productivity," *The Journal of Social Issues,* No. 2, 1956, pp. 43 ff., for a discussion of other studies of this sort.

stockholder. Clearly, the cost of the rights must be taken into account in calculating the entity that is to determine which set of rights is best for the role structure in question.

2. But the type of rights that can be used, their extent, their cost, and their effectiveness, will, of course, depend on the role structure. To each alternative role structure we can associate a best set of rights in the sense of the paragraph above. The optimal set will be determined by choosing the role structure whose associated best set of rights is superior to that associated with any other role structure.

SMALL VERSUS LARGE FIRMS

Does a large firm have any particular advantage over a small firm in setting up a favorable motivation scheme? On many counts, yes. The large firm has ways of creating an inducement scheme to attract personnel that the smaller firm does not. For instance, the large firm can employ a variety of inducements. It is likely to have a more hierarchical structure and therefore can offer greater opportunities for promotion. The variety of roles in the large firm is usually much greater than in the small one and hence the variety of career patterns is also greater. In most cases, the large firm carries greater prestige than the small one. To the extent that prestige of association is a significant inducement the advantage will be with a larger firm. The larger firm is also more visible than the smaller one, so that when a person considers the alternatives open to him, the larger firm is more likely to enter into his field of perception than the smaller one.[9]

The above is not to deny that for some people and some tastes a smaller firm may be able to offer types of inducements not available to the larger firm. For instance, some may feel that in a smaller firm they would have a greater sense of independence about their work, would be less encumbered with red tape, and would have more of a chance of becoming important. This last is the "big fish in the small pond" argument. That is to say, some people might prefer to have a sense of importance in a small organization rather than be lost in a large one, despite the greater opportunities available to them in the larger one.

---

[9] On many of these points see March and Simon, *op. cit.*, pp. 74 ff.

All in all the inducements offered by the large organization are, in most cases, more significant from the point of view of attracting the highest quality of personnel than those offered by the small organization.

In addition, a large firm can duplicate many of the advantages of a small one through decentralization. That is, by breaking up its operations into small plants and small work groups it could create, if it wanted to, many (perhaps all) of the motivational attributes of the small firm. The reverse is not true. The small firm cannot, through organizational rearrangements, create those motivational advantages that the large one has because of its size.

13

The Functioning of the Firm and Organization Equilibrium

OPERATION AND CONTROL

We now look at the operation of the firm from a somewhat larger viewpoint than the one of the individual role-taker.

In any large organization there will be a separation between the major decision-making activities of the direction and control types and direct performance broadly conceived.[1] Connecting these activities will be a host of linkage activities. At one end we visualize a direction-control setup and at the other end a performance setup. Between the two there will be a flow of orders and information in both directions—a circular communication flow, in a sense, which keeps the organization in operation. When that communication flow clogs up, the orderly operations of the organization are interfered with and performance deteriorates. The broad nature of this flow is illustrated in Figure 25. At the top we see those involved in direction roles making various decisions. These decisions must then be communicated to others for performance or action of some sort. Looking at the diagram clockwise we see the linkage activities connecting direction and performance, then connecting the results of performance with the outside world, and finally, completing the circle by connecting both performance information and the outside world with the direction-control center. The decisions are likely to be of three kinds: (1) routine direct performance decisions; (2) expansion, contraction, or other decisions involving changes in the scope or nature of the operations; and (3) control decisions.

---

[1] This distinction must not be confused with our basic notion that every effective decision implies an activity by the decision-maker, although often the act involved may be nothing more than communicating the decision to someone else, who in turn may communicate it to still another person. What we have in mind here is the broad separation of the role structure into primarily direction-control decision-making segments and primarily performance segments.

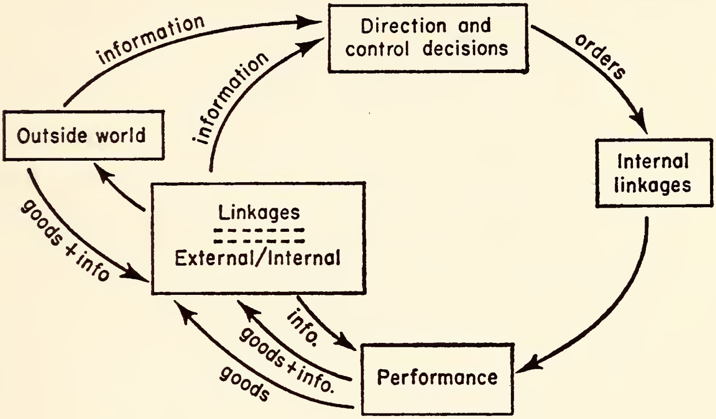


Figure 25. The decision-communication-performance chain.

We may look at the control decisions in terms of some of the role types that are influenced by them. First, a control decision may affect roles themselves. It may direct a man to reinterpret his role or even to abolish his role. Second, the control decision may influence a player. It may cause a man to be fired, replaced, shifted to another post, and so on. Third, the control decision may influence the payoff, by instituting a change in salary, the imposition of a fine, the granting of a bonus, etc. Fourth, the control decision may simply be informative. Somebody may be warned about something, or lauded in some manner, without any specific action. Presumably all control decisions will be made in the interest of maintaining or changing performance in accordance with the desires of the decision-makers.

Next we reach the performance stage. Direct performance involves the actual doing of the things the organization is supposed to do vis-à-vis the outside world, e.g., the production of goods. Indirect performance involves all other activities. The external linkage activities makes possible the relations between the organization and the outside world and connect performance with the outside world. They include selling, purchasing, deliveries, and so on.

The internal linkage activities involve the shifting of goods and materials within the firm and the communication, assembling, and assessment of information relative to the activities of the firm. Thus, internal linkage activities are very much a part of the direction-control-performance nexus. At the performance end, records have to be kept and in some way indices of performance assessed and communicated. Also, the results of external linkage activities must be communicated to the decision-making center. From this point of view we now follow the flow of information from the performance center to the decision-making setup. Once there, this information is utilized to make new decisions and continue the circular flow.

It is not to be imagined that the very rough picture we have painted represents in any way the formalized departments of an organization. This is the very last impression we wish to impart. Rather, we have presented an abstraction of all the interrelated roles and departments within an organization. In actual cases the same department may very well be carrying out direction and performance activities. To some extent there may be an actual separation of major decision-making activities in one department and major performance activities in another. But the real organizational setup will, of course, be complicated by the fact that within each department, even departments of a primarily performance nature, there will also be separate decision-making, communication, and performance complexes.

All that we have attempted to present in this section is a broad operational picture within which the detailed choice sets, activities, and consequences take place. In practice the picture is very much more complicated and contains many more divisions, and interactions between divisions, than the diagram suggests.

ROLE DETERMINATION

We now turn to look at that aspect of operations which determines the roles that make up the organization.

It is generally to be observed that if an organization is viable and durable then many of its roles are more permanent than its players. Or more specifically, the durability of its roles or at the very least of the crucial role types, such as the direction, player-determining, and role-determining roles, must not depend entirely on the duration of the connection of specific players with the organization. The persistence of *all* roles must not be at the mercy of specific players filling them. This is not to suggest that the roles need by any means be rigid. They may, indeed, undergo gradual changes from time to time, or drastic changes may take place at crucial times. But the basic role types must persist if the organization is to persist. At no time must there be doubt as to the role-determining roles of the organization. This point is of obvious importance because we have defined an organization so as to emphasize the role-determining roles, and the dependence of an organization's existence on its role-determining roles. We now have to investigate how roles are determined within an organization.

Broadly put, there are two aspects to role determination: one part is determined by the organization and the other by a player or by a sequence of players.

A role does not exist in a vacuum. It exists within the organization because other roles related to it exist within the organization, and because some role-players, by virtue of the roles they fill, have rights, powers, and duties with respect to certain other roles. Some examples will best indicate what is involved here. Most roles may be said to exist (1) because of a hierarchical structure within the organization and (2) because there is a group or individual having the power and responsibility to act with respect to the filling of a certain role. The simplest case is one in which somebody has the power to hire somebody else to be in a subordinate position to him. The relations between superior and subordinate within an organization are worth examining because other and less clear-cut cases may simply be more complicated varieties of this one.

In the superior-subordinate case the superior may be said to assess the work of the subordinate, to be able to determine his payoff, in part, and to determine which particular individual should fill that role. Where the tasks involved are of a simple routine and very well-established nature, the superior may communicate a complete description of the role. However, such instances are rare. Some characteristics of the role may be communicated, but usually not enough to indicate the detailed execution of the role itself. It is up to the specific role-filler to work out the details and to give a specific interpretation. How this takes place in practice will depend on the interaction between the superior and the subordinate. The subordinate visualizes certain fields of action as part of his role. By experimenting and playing within his field of action he discovers whether the result leads to rewards or penalties, and by judging on this basis he gradually determines what is a workable role interpretation. A sequence of such role-fillers may result in a sequence of role interpretations, some of which may not be too different from the others, but all of which add to the concept of the role in the mind of the superior. It is clear that within this narrow relationship a role relation and role interpretation can be worked out. It is also clear that development of the role itself is determined not only by the superior but also by the history of the various players in their playing of the role.

In a real organization the superior-subordinate relation, and role determination through the superior-subordinate relation, will not be developed in so clear-cut a fashion. Many more people will be involved in the interactions that determine the nature of the role. Yet apart from the number of individuals many aspects will be similar. While there may not be a single superior to hire the role-player, there may be a group or committee, or a series of committees, who will evaluate information and pass judgments that will eventually lead to the hiring of the player. Similarly, the performance of the player will be assessed not by the superior himself, so to speak, but by various individuals carrying out the assessment function. Performance may be assessed directly by one or more people who actually have to assess performance as part of their function. Or performance may be assessed indirectly through the day-to-day relationships between the role-player and those with whom he is related in his role. This type of assessment is especially likely to manifest itself when performance of another man's role, or of other people's roles, is in some way connected with the one under consideration. In a similar way to the superior-subordinate relation some characteristics of the role may be communicated to the player, but the player himself, in his day-to-day activities, will discover and experiment with other possible characteristics of the role. We do not suggest that everything is precisely the same within the large organization as within, say, the primitive master-servant relation. The picture we have just painted should indicate that within the large organization responsibility for a role may be separated from assessment of performance, and from control, so that all of these relations are much more diffuse than is the simple master-servant relation. Nevertheless, the necessary relations will usually exist, and the interaction between the player and other role-takers will gradually determine the nature of the role.

In some organizations player determination may not be made through groups that have a superior-subordinate relation to the player. In other words, the player may be said to be chosen by his organizational equals, or in any event no clear-cut evidence of inequality between the player and those responsible for his choice exists. In many organizations this may be the case with respect to the highest positions. If we wish, we may interpret this situation as being one in which the relation of superior and inferior does exist but is between the player and the group as a whole, perhaps, rather than between any two individuals.

ROLE EQUILIBRIUM AND DISEQUILIBRIUM

In the previous section we have merely indicated the groups that help to determine the nature of a role. It may take a long time for a role to become reasonably stable. In this connection let us examine the conditions for role equilibrium, as well as some of the reasons for role disequilibrium.

A role may be said to be in equilibrium if in the absence of external changes its nature does not change. In this condition the role will have a fixed set of interpretations, and none of the interpretations will result in organization reactions to the interpretation or its consequences such that the player cannot continue with that interpretation. This means that the fields of action in any of the role interpretations in the role are not in a state of flux. The player perceives the alternatives of each field of action clearly, and the nature of the play is such that no new experiences can be perceived by any possible choice set. Stated differently, the consequences of any of the possible choice sets will not reveal either new alternatives, new fields of action, or new desirable role interpretations. This is one half of the equilibrium picture.

The other half is that no player chosen by the organization for the role could through his choice sets discover new alternatives, fields of action, or permissible role interpretations. The implication is that the organization can distinguish consequences that are due to the nature of the role as against consequences that are due to the way the role is played. In other words, if undesirable consequences occur, and if the role is in equilibrium, these consequences will be attributed to the player and will set in motion the search for a new player, rather than being attributed to the role and instituting an attempt to change the role.

Some causes of role disequilibrium. A role may be in disequilibrium for at least one of the following reasons: The role may be evolving (that is, not fully determined), roles may be in conflict, players may be in conflict in non-fully determined roles, or the role structure may be in disequilibrium.

Where the role is still evolving there usually will not be a clear knowledge of the alternatives, and the player will still be attempting to explore some of the alternative sets. On the organization side, there may be a lack of knowledge of the capacities of the average player, and as a consequence when the organization is dissatisfied with the nature of a play, it may not be able to determine whether the role or the player is at fault. Even though the player, through his exploratory activities, may have found some acceptable role interpretation, he may nevertheless seek other role interpretations in the belief that they may, in some sense or other, turn out to be superior to the one he has chosen. Also, there may be a lack of understanding between the player and the organization, or parts thereof, as to what is an acceptable role interpretation. Through experience such understanding may gradually evolve and the role approach an equilibrium state.

A role previously in equilibrium may be sent into disequilibrium by the choice of an unusual player. In other words, a new player may discover role interpretations by-passed by previous players but also acceptable to (or found superior by) the organization.

One of the major causes of role disequilibrium is role conflict. The two roles in conflict may be roles played by the same person or roles played by different people. Here we may introduce the concept of dominating and inferior roles. A dominating role is one that takes precedence over another role when it comes into conflict with it in the sense that its interpretations remain intact while those of the other role change. An inferior role is one that is reduced in some sense when it comes into collision with another role. A role may be said to be reduced if the player is forced to choose an interpretation with fewer fields of action in it than were contained in the interpretation given up. In the usual case domination and inferiority are to be seen as relations between roles in disequilibrium, or at the very least where one role is in disequilibrium.[2]

---

[2] It is conceivable that role A dominates role B but the role structure is such that the two role-players do not have any contact with each other and

A role may change because of role conflicts within a player as well as role conflicts between players. We have already argued that an unusual player may upset a role in equilibrium and that almost any player can help to form a role not yet in an equilibrium state. From this point of view role conflicts within a player may determine, depending on the resolution of the conflict, the nature of the role. If a nonorganizational role dominates an organizational role, and if it does so for all likely players, then the organizational role will have to be reduced. For example, if a man's role in the political life of his community dominates his work role, and if he is an unusual and valued player in the latter role, his work organization may permit him to reduce his work role. The same may hold with respect to possible conflicts between family and occupational roles.

All we intend to suggest here is that role conflicts within an individual can change the nature of a role and be a cause of role disequilibrium within an organization. Such role conflicts may occur because (1) the two roles compete for the player's time, (2) performance results in the pursuit of conflicting objectives, or (3) there is a change in the life cycle of the average player so that roles which were previously not in conflict now are in conflict.

Probably the more significant conflicts within an organization involve role conflicts between individuals rather than within an individual. We have in mind here not the conflicts that would arise because of the special personality of an unusual individual but those that would arise if average players were appointed to the roles in question. Some reasons for such conflicts are as follows: The two roles may conflict because they contain the same fields of action and there is uncertainty over whose responsibility it is to play on the conflicting field of action. This may lead to conflicting choice

---

hence actual conflict is prevented. In other words these roles are potentially conflicting but as things work out the conflict does not manifest itself. This can happen when the flow of information is imperfect, or when it is withheld from the player in the dominant role. E.g., a foreman may continue to hire some workmen despite the fact that the personnel manager has jurisdiction in this area, as long as the personnel manager does not find out what the foreman is doing.

sets. Conflicts may also arise because roles are "too broadly conceived" from the organizational point of view. Although some role interpretations within the two roles do not lead to any conflict, because of the breadth of the roles some players may choose interpretations that do lead to conflict. Also, though the roles themselves may not be in actual conflict, some of their interpretations could be *played* in such a way that from the organizational point of view it looks as if conflicting objectives are being pursued.

Finally, role disequilibrium may exist because the role structure in which the role is embedded is in disequilibrium. That is to say, the role in question would remain unchanged if the other roles stayed put, but since the other roles themselves are changing the role in question is forced to change accordingly.

ROLE STRUCTURE EQUILIBRIUM AND DISEQUILIBRIUM

Role structure equilibrium can be looked at from the point of view of the web of role relations.

In an organization no role exists alone. This follows from our definition of an organization as a set of related roles that have certain characteristics.[3] Thus, each role is connected in some way with at least one other role and often with a host of other roles within the organization. The connections manifest themselves through the alternatives in the fields of action of one role that impinge on the alternatives available to other roles. Thus, alternative A in the field of action Alpha may affect the alternatives open in field of action Beta, which in turn may affect those open in a third field of action. In a similar way the alternatives in other fields of action impinge on the alternatives in the fields of action under consideration. For example, if a personnel manager offers a new post to someone within the organization, this act impinges on or changes the alternatives of that person within the organization, and his decision will in turn change the alternatives open to other individuals. A *routine*

[3] A role that was not connected with a role-determining role (or with either a payoff or a direction role) would not according to our definition be part of an *independent* organization.

alternative in a routine field of action may bring up a contingent field of action in another role, or routine alernatives in one field of action may simply impinge on the routine alternatives of another field of action. An example of the latter would be the processing of an order between several departments. If in one department a decision is made with respect to what is to be produced, then someone in the production department may have to decide precisely how it is to be produced.

Most activities in an organization result from the choices made within the network of connections between the routine fields of action within the roles that make up the organization. In other words, the activities are the observed consequences of the interrelated decisions that take place almost continuously within a firm. Since there is some choice and discretion given to various players within the organization, and since the environment which, in part, influences decisions will change from time to time, there is no need to expect the actual activities that take place to be the same period after period. The vast number of choices made will result in a changing pattern of activities from time to time. But this pattern will not be a random one. In a well-run organization, if some decisions lead to activities that seem "out of line" with policies or objectives, there are usually control activities to generate counteracting activities so that the net result is within certain tolerances. Needless to say, at times real difficulties will arise—bottlenecks will develop, or inconsistent decisions will be made by different groups within the organization, or significant errors will occur. Under such circumstances the contingent fields of action in some of the roles will come into play to attempt to counteract the difficulties. In other words, we visualize the operations of the organization in terms of a pulsating set of activities which are a consequence of the vast number of interconnected detailed decisions made within the web of related fields of action.

Normally, there is some logic to the relatedness of the fields of action, the decisions made, the activities performed, and the objectives of the organization as interpreted by the control group within the organization. Usually, it is impossible to determine whether the activities exactly fit the objectives of the group. As a consequence some tolerances must be permitted. When these tolerances are clearly violated, attempts will be made to correct the violations by additional decisions, or at the very least additional decisions will be made in an attempt to prevent their recurrence. In some cases the prevention of a recurrence will necessitate a change in some of the players. But it can readily be seen that activities may vary quite a bit without necessarily implying that such variations call for a change in any specific role. Therefore, if the role structure is in equilibrium to begin with, it may be consistent with a large number of choice sets, but of course not all. Some choice sets will conceivably lead to outcomes that fall so far short of organizational anticipations that they engender investigations ultimately leading to role changes. Thus, it is the interaction between elements of the web of role relations—the usual outcomes that occur as a consequence of such relations—as compared with organizational anticipations, that determines whether or not the role structure is in equilibrium.

A role structure may be said to be in equilibrium when all permissible interpretations of roles by representative players, and all possible choice sets by such average players, do not lead to any change in any role within the structure. Under these circumstances, of course, the organization may be dissatisfied with either a move, a play, or the general performance of a player, but it does not seek as a remedy to this situation any change in the role structure as such.

It is not unusual to find that role structures remain fairly constant over relatively long periods of time. Looking at a reasonably well-functioning organization whose scale of operations is relatively stable we are unlikely to find its role structure changing rapidly. It is of interest to speculate on why some role structures should have a high degree of constancy.

The answer must lie, at least in part, in the fact that effective performance usually requires a great deal of constancy in the role relationships. Day-to-day internal performance often involves much role interdependence. And this suggests the desirability of a lack of change in role relationships if performance is not to be interrupted unduly. If performance depends on more than one role, things are obviously more difficult. The role-player then has to do his job and *at the same time* try to discover the connection and relation between his performance and that of the next person. Within a fairly constant role structure these connections are well known, often taken for granted, and the player can concentrate on performance rather than spend part of his time attempting to discover role connections.

The value and efficiency of well-understood role connections can be readily seen if we examine some types of role relations. For example, consider unidirectional role relations. Suppose that a person in role A has to assess certain information and communicate the results to someone in role B. If the one in role A does not know the connecting role—that is, which is role B—then clearly the flow of information through the organization will be considerably hindered. Even if there is such knowledge, if these relations change very rapidly, then obviously a great deal of time will have to be spent on communicating and learning the changes in roles and role relations rather than on performing the routine activities. A similar argument, but an even stronger one, can be made with respect to coöperative role relations. The crux of the matter is simply that if the role structure changes rapidly, it becomes very costly to stop at each point in one's activities in order to question the appropriate procedures and determine the right role relations. We may observe in passing that the same would not be true for fairly rapid changes in personnel. Person A need not know that he is responsible to a specific person B. It is sufficient for him to know that he is responsible to a person in a specific role such as, let us say, the general manager in charge of something or other.

The above arguments can readily be seen to be equally true with respect to the external linkage activities of the organization. Those outside of the organization who deal with it have to know with whom they are to deal. Once again it may not be necessary for them to know specific individuals but it is helpful if they at least know the appropriate posts or roles involved. For example, the external linkage activities would become exceedingly difficult to handle if the outsider has on one occasion to deal with the credit manager on a certain type of problem, and on another occasion with the vice-president in charge of production, and on a third with some other position for the very same problem. The relations between the organization and the outside world have to be carried on usually through contacts between individuals in fairly well-recognized roles. If the roles change too frequently, such contacts are impeded. If they remain fairly constant, the ease of making such contacts obviously improves.

EQUILIBRIUM ROLE STRUCTURES AND OPTIMAL ROLE STRUCTURES

The argument we wish to develop in this section is that the equilibrium role structure of an organization need not be an optimal structure. First we must elaborate on what we can possibly mean by the concept of an optimal role structure.

The word "optimality" suggests a criterion for ordering different situations. But the question always arises: What criterion should we adopt? With respect to an organization, from whose point of view is the role structure to be deemed optimal? One might argue that it should be optimal in the sense that the understood objectives of the organization should be met as much as possible. But this presumes a state of affairs that need not exist. That is to say, it presumes that the objectives are well understood. We will examine in a later chapter to what extent an organization can be supposed to have specific objectives. For the time being let us have in mind the objectives of those in the directing role complex, and furthermore, let us assume that among those in this role complex uniform objectives exist. As we proceed with the argument we may have to relax this last assumption, but until we do so specifically we assume a uniformity of viewpoint within the directing group.

Suppose that the existing role structure is an equilibrium one. Is it optimal? The following are some of the main reasons why it may not be optimal.

1. The role structure may not be optimal because of a lack of knowledge or experience. Previous role-players may not have experimented sufficiently with role interpretations, so that too narrow a view of the role has developed. Since these interpretations produced reasonable, though not optimal, results, there was no incentive to change the role structure. The point of the matter is that one may not know whether the role structure is the best possible. All that one may know is that the existing role structure is as good as or superior to those tried previously, but there is no way of comparing it with those that have not been tried.

2. A role structure may not be changed because of the fear of a disequilibrium situation. In a complex organization the relations between roles are likely to be exceedingly important. As long as things seem to be working reasonably well at the connecting points, why try to supplant the role structure with an alternative, even if the alternative would very probably be a superior role structure? For while the alternative role structure may be superior it will not necessarily be an equilibrium role structure. That is to say, the consequences of the choice sets may be such that further changes in the role structure will be attempted, which in turn may lead to still further changes, without any specific knowledge as to when this period of change might end or *what* might result at the end. The fear involved is not with respect to the initial change-over but with respect to the possibility that the initial one may lead to other change-overs whose value may not be perceived or perceivable.

3. Even where the change-over is expected to be an equilibrium one, the cost of a change-over may be very high. A new role structure involves a new set of relations between roles and therefore a costly rearrangement of everyone's duties, connections, and responsibilities. Not only would many people have to become used to doing different things but they also would have to become used to different relations with others. This is not to suggest that a more desirable role structure *need* be too costly to introduce, but it *may* be. Thus an existing role structure may persist even though it is known not to be optimal.

4. A nonoptimal role structure may persist even in the face of

knowledge of superior role structures if the groups in authority are divided on objectives or on their assessment of the desirability of a different role structure.[4] It is not at all inconceivable that those in authority may not have uniform aims and objectives, or uniform views as to how best to pursue aims and objectives. As a result the existing role structure may be inferior from the viewpoint of both groups to their own alternatives, but they cannot agree on the superiority of any specific alternative. Furthermore, the new structure may involve placing some people in superior positions and others in inferior positions to those they now possess. The ones who believe they may lose by a change-over may be the dominant group and as a consequence the existing role structure will persist. Entering here, clearly, are the power relationships and attitudes toward risk and change of those who are important in the organization.

CONCLUSIONS

1. Roles are determined by the interplay of three forces: by players in their choice of role interpretations, by the history of plays and the organizational reactions to such plays, and by the organization—which in reality manifests itself through the web of role relations. In a stable role structure roles do not depend on the participation of specific players. In a viable organization specific roles may change in detail, but the basic role types (such as the player-determining and role-determining types) must continue to exist, and their existence must not depend on the participation of specific players.

2. We distinguish between the equilibrium of a role and the equilibrium of a role structure. A role is in equilibrium if in the absence of changes external to the role it does not itself change. This means that the persistence of the role does not depend on the way it is interpreted and played. In turn this implies that the organization can distinguish between consequences attributable to poor playing and those resulting from the nature of the role.

---

[4] Here we obviously relax our assumption that uniform objectives for the organization are held by different members of the control group.

3. A role may be in disequilibrium because (a) the role is evolving, i.e., it is not yet fully determined, (b) the role-player is in conflict with others and the role is not yet fully determined, (c) the role is in conflict with one or more other roles, or (d) the role structure is in disequilibrium.

4. A role structure is in equilibrium when all permissible interpretations of the roles within it by representative players, and all possible choice sets made by such players, do not lead to any changes in any of the roles within the structure. The interaction between elements of the web of role relations and the usual outcomes of such relations, as compared with organizational anticipations, determines whether or not the role structure is in equilibrium.

5. Actual role structures often have a high degree of stability or change very slowly. This is partly because clearly recognized and well-understood role connections, which are fostered and reinforced by a stable and enduring role structure, facilitate the flow of information, the assessment of information, and the smooth functioning of the direction-communication-performance chain. In addition, the effective operation of the external linkages depends on clearly recognized roles, which in turn are influenced by the stability of the role structure.

6. An equilibrium role structure need not be an optimal role structure. (a) An organization may settle on a role structure that is satisfactory but not optimal because the organization has not experimented sufficiently to learn that a better role structure exists for it. (b) Since role structures are complex entities, and since there are advantages to stability in the role structure, those in control of the organization may believe that the cost of experimenting with changes in the role structure in the hope of discovering a superior structure is not worth the effort. (c) A shift to a known superior role structure may not be attempted because of the fear of the disequilibrating effects that such a shift might generate. (d) Role structures are very difficult to evaluate, and there may be disagreement within the organization as to whether an alternative to the existing role structure is superior, although in fact the alternative

may be in the light of the recognized organizational objectives. (e) Some members of an organization may be made worse off, or believe that they would be worse off, given their personal goals, by a shift to a superior role structure. Those who believe they would lose by the shift may be in a position to prevent the change.

14

Some Economic Aspects of Organization

SPECIALIZATION AND ADMINISTRATION— A BRIEF REVIEW

We may look upon the question of the economics of organization from two points of view. First, are complex economic organizations economical, and if so why? Second, what are the characteristics of an efficient organizational structure?[1] In this chapter we shall examine the first of these questions. Our answer will be that while in some cases large complex organizations are not efficient, there are no a priori reasons to believe that they cannot be, at least in some industries and under many circumstances.

The economic desirability of a complex organization is closely related to the sort of ideas we considered in the first half of this book—namely, the relations between specialization and efficiency. A complex organization involves the coördinated specialization of activities. The organizational element enters the picture not through

[1] See Chapter 15.

the fact of specialization itself but through the fact that specialization makes coördination necessary and coördination is achieved through organization. Also, organization (the administrative sector) itself may be more or less efficient, and specialization, its advantages, and its limits play their part within the administrative sector.

Let us look at specialization—that is, the separation of activity segments into smaller segments and the restriction of players to these smaller segments—and, noticing the location of the cuts between activity segments, consider why certain types of cuts between activities should be economical. Throughout we shall emphasize specialization within the administrative sector.

Up to a point specialization is increased and economies are obtained by the following means: (1) by separating the direction-decision activities from direct performance activities;[2] (2) by dividing direct performance activities into small activity bundles; (3) by separating routine decision-making from nonroutine decision-making; (4) by separating control of performance from direct performance, and also by separating the measurement of performance from direct performance; and (5) by separating the transformation of performance into payoffs from direct performance itself.

Consider for a moment the separation of routine decision-making from nonroutine decision-making. Where possible it is often desirable to substitute choice rules for individual discretion. Also, it is normally efficient to routinize the decision-making-direction-performance mechanism as much as possible. And finally, internal linkage and external linkage activities may well be routinized as much as possible. What is left for other roles will be the nonroutine and contingent decision-making activities in these areas. The other side of the coin of routinizing activities is to enable the organization through specialization to carry out explorative, creative, and expansional decisions and activities without interrupting routine production.

---

[2] In accordance with our overall approach we assume as usual that those in direct performance roles do make decisions as seen from a strictly formal viewpoint, although from a practical viewpoint they may often be exceedingly trivial decisions, e.g., whether or not to follow a certain standing order. They are trivial decisions because they are predictable in the vast majority of cases. The worker will usually obey the order. Nevertheless, alternatives, choice, and a decision-making process can be said to exist.

In general, essential economies of production may be said to arise from the fact that the direct performance activities are carried out more efficiently through specialization. All other activities may be looked upon as necessary for production but not directly involved in the transformation of raw materials into goods or services. However, it must not be supposed that under unspecialized production indirect activities are not involved. This may sometimes appear to be the case simply because the individual carrying out production activities is simultaneously carrying out indirect performance activities. To some extent such indirect activities may be truly costless in the unspecialized state if the producer can carry them out without reducing efficiency in performing the direct activities. At the same time it should be kept in mind that this is likely to be true for only some, and probably for an exceedingly small portion, of the indirect activities. On the other hand, the specialized mode of production usually requires additional activities which are not required in the unspecialized situation. The separation into convenient bundles of those activities that would have to be carried out in any event in the unspecialized state can lead to increased output because of both the allocation and the concentration advantages of specialization.[3] But we must debit against such increases in output (1) the additional cost of those activities that are carried out simultaneously with other activities in the unspecialized mode (and are therefore costless) but cannot be so carried out under more specialized production, and (2) the cost of the additional and/or new activities that are required by the increase in the degree of specialization.

To see what is involved here let us briefly consider how we might gradually increase the degree of specialization in the production of a commodity *beginning with a completely unspecialized mode of production*, taking into account as we proceed the organizational requirements of specialized production.

---

[3] See Chapter 8.

1. The direct performance activities are separated into bundles and assigned in such a way that no individual carries out all of the activities necessary for production. The greater the degree of specialization, the greater the number of segments, with persons assigned to them, into which these activities will be separated. The mere fact that they are separated creates the necessity of establishing means of coördination. This is the beginning of the organizational problem.

2. The minimum amount of organization required is the establishment of a direction apparatus and a linkage apparatus. Since people are each performing a segment of the activities, they have to be assigned to these activities and the assignment must be communicated. The linkage activities will involve communication between direction and performance, and the necessary transportation of materials and men between the segments of direct performance. With respect to the former, a two-way flow is required. Orders must be carried from the direction to the performance sector, and performance information must be conveyed back to the direction sector.

The direction sector obviously involves activities that to some extent have to be carried on in the unspecialized mode of production. But they may be performed more effectively in the specialized mode because of the allocation and concentration advantages. The same may be said for the linkage activities.

3. In order to keep the organization functioning, external linkages to the outside world have to be established. It is these external linkages—for example, the obtaining of orders from customers and the ordering of raw materials—that make possible the continued operation of the organization, determine in part the payoff of the organization, and decide, again in part, the success or failure of the organization as seen from the viewpoint of the outside world.

4. Now the rest of the organizational structure has to be filled out. The following activities are necessary for the perpetuation of the organization although they are not strictly required for day-to-day activities in the short run. In other words, we can conceive of day-to-day activities proceeding for some period of time without having the following role types come into play.

Certain minimal control activities must be introduced. The necessity arises for two reasons. First, the quantity and quality of production has to be controlled. At the very least, data with respect to quantity and quality have to be gathered, analyzed, and communicated to the direction sector. Also, the direction sector must in turn communicate new directives whose aim is to correct flaws in production methods, quantities, quality, etc. Second, control activities have to be introduced in order to make sure that individuals in fact carry out the directives or intentions of the direction sector. Not all members of an organization can be assumed to be motivated by the same considerations as the directing group, or even by the presumed organizational objectives. Third, since an organization involves many players and we assume that players may leave at will, and since they are not immortal, some means has to be introduced for player selection and player succession. Fourth, since roles are not predetermined by definition, a scheme for role determination as well as for the reconsideration of roles must be set up. Fifth and finally, the payoff to the organization has to be distributed among the members of the organization and therefore a means of payoff determination has to be included.

All of the organizational elements just mentioned are not independent of one another. For instance, player dismissal, a part of the player-selection mechanism, may very well be part of the control scheme. A system of bonuses and penalties, which may form a part of the payoff-determination mechanism, may also be part of the control scheme. Promotions within the firm may be tied to the payoff mechanism. Finally, the role-determination scheme is all-pervasive since all of the activities are connected with the system of roles that exist. Indeed, it is these interrelations that make the collection of roles an organization.

ORGANIZATIONAL ECONOMIES AND SCALE OF OPERATIONS

Many of the activities connected with production have counterparts in unspecialized production and almost all of them can, to some extent, be carried out more efficiently through specialization within an organization than in the unspecialized state. Here we shall examine why this will be the case at least up to a point, emphasizing as we proceed some of the organizational elements of the phenomena involved.

1. We have already discussed at length (Chapter 8) how the direct performance activities can be carried out more efficiently through specialization. We need not go into any detail on this point.

2. Let us begin by considering the juncture between the organization and the world at large, i.e., the external linkage activities. In the unspecialized case the individual will usually have to stop his performance activities in order to engage in linkage activities. Now the nature of many external linkage activities is such that they are especially time-consuming if carried out on a small scale. For example, the individual producer has to find customers for his wares. He may have to search out among a very large number of people that particular person who is interested in his own productive efforts. In the usual discussion of barter it is assumed that the people involved in an exchange easily find each other. But actually an unusual amount of effort may be required. For example, in backward economies people who are partially self-employed in agriculture often spend large amounts of time traveling considerable distances to neighborhood markets in order to exchange items of exceedingly small value.[4] Clearly, to the extent that a large number of potential buyers and sellers have to be canvassed in order to effectuate an exchange this activity can be done more efficiently through a large organization at a large scale of operations. Of course, some people may be producing and selling highly standardized commodities, such as staple agricultural products, and markets for these may be so well organized that an individual's external linkage activities are minimal compared to his direct performance activities. But note that under these circumstances an organized effort must exist somewhere. Somebody or some group must somehow provide the place and the exchange machinery. The provision of such market

---

[4] See P. T. Bauer and B. S. Yamey, "Economic Progress and Occupational Distribution," *Economic Journal*, December, 1951.

machinery usually involves an organizational rather than an individual activity.

3. Closely related to the external linkage activities—indeed to some extent a part of them—are those exploratory and related activities of an organization that are necessary if the organization is to grow. Under individual unspecialized production the producer usually finds it too costly to take time off from regular production in order to engage in those uncertain exploratory activities necessary for expansion, growth, and change. Also such exploratory activities are often disproportionately time-consuming compared to the production activities and as a consequence may not be efficiently undertaken by a single individual. Again, a man who is good at producing goods is not necessarily good at the exploration of new markets or other economic opportunities. Thus, individual production may have to be carried out without the means to take advantage of improvement possibilities, whereas an organization can have built-in mechanisms for seeking out and taking advantage of growth opportunities.

4. The direction role may be carried out more effectively, up to a point, within the organizational context, where it can be specialized to a considerable degree. Direction is, after all, an activity that must exist both under individual production and under an organizational setup. Consequently the usual specialization advantages—the allocative, concentration, and qualitative advantages—can manifest themselves under organized production.

5. As suggested earlier, there are two aspects to control, a production aspect and an organization aspect. To some extent the production aspect of control is required under individual production. The individual has to determine whether his output, at least qualitatively, matches market requirements. However, this is often a trivial aspect of his activities, and he may be able to carry them out simultaneously with other activities. Nevertheless, where production is for an undetermined market, especially qualitatively, the large organization through specialization may carry out even this control aspect more efficiently.

On the other hand, the organizational aspect of production is something that does not exist for the individual, since there is no separation between individual and organizational motivation and hence no requirement for control on this basis. But even here organizational production has an advantage for certain types of people. Some individuals are so lacking in self-discipline that they produce more efficiently under the discipline that an organization can provide. Of course, a very small organization may provide the necessary discipline, and hence this is not an advantage that will necessarily increase with scale of organization.

6. The individual producer is not concerned with personnel selection as a day-to-day activity, but something closely akin to it is not entirely absent from his scheme of productive life. At some points he must make decisions as to the types of economic activities he will engage in, and these may be similar to the decisions an organization has to make with respect to the assignment of activities to an individual. Note the close connection with the problem of role determination. An individual producer has to assign to himself a sequence of activities. But this job may be carried out more effectively if an outside organization helps in determining the sequence. On the other hand, the outside organization cannot take into account all of the information that an individual may have about himself. In sum, though, the problem of role determination would appear to be much more complex for an organization, since every time one role is changed a host of other roles may be affected and perhaps changed accordingly. Furthermore, these interrelations may grow more than proportionately to the growth of the organization itself, since a given increase in organization size can easily result in a more than proportionate increase in the relations (and necessary contacts) between different roles.

7. Payoff distribution as a problem does not exist for the individual; hence this organizational cost does not have its counterpart in individual production.

We have seen that a great many of the activities that have to be carried on under individual production have their counterparts in organizational types of production. Since they can be carried out in a specialized manner within the organizational context, the advantages of specialization achieved thereby will accrue to the organization. In other words, the potential advantages of specialization are as pertinent to the administrative end of things as they are to the strictly productive end. Therefore, there is reason to believe that most activities (indeed it is difficult to think of significant exceptions, even in the cases of personal services) can be carried out to advantage, at least up to a point, through organizational means. But such advantage may be offset by those activities required within an organization that are unnecessary under individual production.

We have seen that increases in specialization and scale of operations can be associated with direct production economies and organizational economies. It is also possible, as we shall argue in this section, and as is well known generally, that beyond some point increases in degree of specialization, in scale of operations, or in both can lead to organizational diseconomies of various types. But at the same time that diseconomies are occurring in some areas of an organization, further organizational economies may manifest themselves in other sectors as the firm expands. It is clear that some organizational diseconomies do arise with increases in scale, at least beyond some point, and that they become—again beyond some point, and in some industries—more significant than the economies associated with increases in scale.[5] It is such possibilities that we now discuss.

ORGANIZATIONAL DISECONOMIES

In Chapter 8 we discussed the limits to the degree of specialization that arise in connection with the indivisibility of certain factors, or the fixed size of certain other factors. We repeat here in summary fashion some of the main points. Individuals as well as many activities come in fixed sizes and capacities. When something is fixed in size or capacity there are two aspects that determine efficiency. On the one hand we must consider the possibility of underutilization, and on the other the possibility of overloading.

---

[5] It is not possible to say as we consider larger and larger firms that the balance *must* always shift in favor of the diseconomies of scale. (We consider this in the next section.) But even if this were the case, it would not preclude the possibility of a firm's being so large, before the balance shifts in favor of diseconomies, as to be able to supply the entire market at its lowest cost per unit of output. This is the Achilles heel of the theory of pure competition. See J. S. Bain, *Barriers to New Competition*, Harvard University Press, Cambridge, Mass., 1956, especially pp. 61 ff., on this and related matters.

The fixed size of an individual sets limits to the degree of specialization. Activities can become so minute that they are uneconomical, or more difficult to perform than bundles composed of larger activities. For example, a certain motion may be required to pull a switch. But it cannot be effectively specialized further by breaking it up into two motions. At the same time, breaking up the activity bundles into too small units may leave unused capacity within the individual. For example, an individual carrying out activity A may be able to carry out activity B simultaneously at no extra cost. In some cases the complementariness of activities may be of such a nature, or the results of the two activities have to mesh in such a way, that it is more effective to have these two activities carried out by the same person rather than by different people. Before any of these limitations to further specialization become serious, however, the organization may already be a very large-scale one.

In addition it is of interest to note that limits to the degree of specialization at the direct performance level do not necessarily set the limits to firm size. There may always be a larger size firm at which it pays to introduce a specialized activity, at some level in a hierarchical organizational structure, that it did not pay to introduce at the smaller scale of operations. To see this we consider briefly the relation between the hierarchical structure of an organization and the utilization of specialized services.

Hierarchical structures make possible enormous increases in firm size with only small increases in the number of hierarchical levels. Suppose that the "span of responsibility" is 10 to 1, say 10 "supervisees" to every supervisor, or 10 directees to every director. By the span of responsibility we have in mind a portmanteau concept to cover such notions as the span of command, the span of control, the span of supervision, and so on. A five-level hierarchy could direct a firm of 100,000 people at the lowest level, and a six-level hierarchy could direct a firm of one million people at the lowest level, and so on. There would be 100,000 directors at the first echelon, 10,000 at the second, and so on. Now, it may be economic to introduce a specialized activity that would service the 100 directors at the fourth level, whereas it would not pay to have this specialized activity with the smaller-scale firm having only 10 directors at the fourth level. In other words, the advantage of introducing specialized activities may continue almost indefinitely if we consider introducing them to service administrators at successively higher hierarchical levels. Of course, whether the advantages of such specialized services are greater than the costs of increasing the proportion in the administrative group that results from a lengthening of the hierarchical chain is an empirical matter which will depend on the specific facts of the situation.

It is nevertheless true that there are many circumstances in which beyond some point larger firms will be administered less efficiently than smaller ones.[6] To see why this might be the case we will try to show that, if the span of responsibility is assumed to be more or less fixed, the hierarchy will grow in greater proportion than the scale of operations, and in consequence, beyond some point the increased administrative costs arising because of the more than proportionately growing hierarchy can be more significant than other advantages that accrue to the firm as a result of the expanding scale of operations.

The increased hierarchical costs will at some point offset the advantages of specialization at the bottom level of the hierarchy. They may also eventually offset the advantages of specialization at the higher administrative levels. In addition, the increased number of potential interactions between individuals at the same level that can occur as a result of scale expansion may also raise administrative costs. But our analysis at this stage, we repeat, will be based on the notion that the span of responsibility remains more or less fixed with changes in scale. (We relax this assumption later in the chapter.)

---

[6] Firms will often not operate as efficiently as they possibly could for their size. Large firms have more of an opportunity to develop complex organizational structures—not all of which are optimal, and some of which may involve unnecessary controls and other bureaucratic inefficiencies. A small firm may, to some extent, be sheltered from such follies by its very smallness, which may the more easily enable it to perceive the lack of desirability of a highly complex form of organization.

1. *Direction costs, specialization and scale.* The basic argument depends on the connections between the span of responsibility of an individual, the length of the hierarchical chain within an organization, and its relation to the degree of specialization. We want to establish that beyond some point as the degree of specialization is increased it is necessary to increase the direction hierarchy within the organization. To do so we introduce the concept of the basic work team—the members of the minimum size group that carry out *one complete set* of the direct performance activities of the organization for a given pattern of specialization. There may be more than one such team since one way of increasing the scale of operation is to duplicate the number of basic work teams. Thus by a basic work team we mean the group that carries out one set of all the direct performance operations; e.g., the nonadministrative workers in a plant in a company that possesses a number of identical plants could be a basic work team.[7]

Now, as the degree of specialization is increased, the size of the basic work team is also increased. This follows directly from the definition of increased specialization—namely, an increase in the division of the activities among workers.

Someone has to be responsible for the activities of the basic work team. It may be one person or a large number, depending on the size of the work team and on organizational considerations. These people may be called foremen, section heads, and so on. At the very least the members of a basic work team must be able to go to someone in cases of conflicts, unforeseen contingencies, or other situations, depending on the nature of the work. In addition, directions have to be issued to the basic work group.

Now, our main assumption is that the span of responsibility of any one person is limited. Also, the larger the basic work group, the greater the likelihood that contingencies and conflicts of one type or another may occur. And further, the larger the work group, the greater the connections between members of the group and therefore the greater the aspects of the production process that may have to be coördinated. Since one individual's activities may be connected with more than one other individual's activities, we can expect the increase in such connections to be greater than the increase in the size of the group. The greater the size of the group, therefore, the more important a mistake may be, since it will affect a larger number of individuals' activities.

---

[7] The basic work team is not restricted to the plant. It could cover a number of plants if the activities in different plants involved different aspects of the production sequence; e.g., one plant may produce engines, another body parts, and so on.

For the time being let us suppose a fixed span of responsibility for everyone in a supervisory (i.e., responsible for others) capacity.[8] Let us call each individual a first-echelon director. The greater the number of individuals in a basic work group, the more first-echelon directors required. If first-echelon directors are in a fixed ratio to members of the work group, they will increase in proportion to the size of the work group. But since the first-echelon directors are responsible for different segments of the activity spectrum, their work must also be coördinated. Hence, for some number of first-echelon directors there must be one or more second-echelon directors and so on up the line. Clearly, specialization requires a hierarchical organization.

Furthermore, the greater the degree of specialization, the more steps in the hierarchy, *again if the span of responsibility remains constant*. It follows that the sum total of all directors must increase more than proportionately to the size of the basic work group. Thus, direction costs per man in the basic work group will rise with the extension of the degree of specialization beyond some point. And if there are diminishing returns (beyond some point) to the further extension of specialization, then the rising direction costs per man will operate in the direction of limiting the further extension of the degree of specialization at the basic work group level.

With respect to scale, a similar argument holds. As the number of basic work units is increased, these work units themselves have

---

[8] Later in this chapter we drop the assumption of a constant span of responsibility.

H B

to be coördinated, and their first-echelon directors have to be responsible to second-echelon directors, and so on. Thus scale itself, apart from increases in specialization, will increase the extent of hierarchy and eventually the administrative costs of organization—the span of responsibility still assumed to be constant.

2. *Control costs and degree of specialization.* In an ideal world in which everything is foreseen, all directions are carried out precisely as expected, and every subordinate pursues the same objectives as the director or directorate, control as such will not be necessary. But in the actual world attempts to control operations are made for at least two reasons: (a) to minimize the deviations between actual performance and some expected standard of performance and (b) to minimize the deviations of attempts to pursue organizational objectives as against the objectives of the players. The last may be further broken down into two categories, those in which the player's objectives are legitimate within the formal rules of the organization although they do not maximize performance (such as pursuing a career objective that may conflict with the organizational objective) and those that are outside the rules of the organization or even outside the rules of society—dishonesty, disloyalty, etc.

Given the need for control, we now examine some of the factors which determine the extent of control and how these are related to increases in specialization and scale. The degree of control necessary will depend upon a number of interrelated factors, prominent among which are the length of the chain of production, the degree of hierarchy, and the diffusion of responsibility and accountability within the organization. We have seen that specialization increases the chain of production and the degree of hierarchy, and consequently the degree of control. If the gains from further specialization decline and if the control costs are not a decreasing function of specialization (indeed, they are likely to be an increasing function), then we may expect a limit to be set by increasing control costs to the extension of specialization. Let us look at these matters in slightly greater detail.

Specialization increases the number of necessary linkages between performers and as a result it increases the consequences of errors which may occur. This is especially true if, as we have argued, production functions are partly sequential. Therefore, greater control costs are necessary as specialization increases.

The number of linkages required in the production process will increase at least at the same rate as the increase in specialization and probably at a greater rate. If control costs do not have any scale advantages—and beyond some point control techniques can probably not gain from further specialization—then it follows that control costs will, beyond some point, increase with the degree of specialization.

In addition, the greater the chain of production,[9] the greater the degree of hierarchy, the more important certain decision-making roles become, and the greater the risk of a role-player's not following organizational objectives.

With increased growth, complexity, and greater hierarchy will often go a diffusion of responsibility and a decrease in the assessability and accountability of the players of a role. Role-players may find, under these circumstances, increased opportunities for not following organizational objectives. In such cases career objectives may conflict with organizational objectives. For example, someone in a responsible position may be supposed to train a subordinate for that position, but in order not to increase the number of competitors he may find it expedient not to pursue this part of his job very seriously.

Finally, there is the matter of what may be called communication noise. As the number of linkages increase, and as directives have to be translated into more specific directives at a greater number of junctures in the chain, the translations become less and less reliable versions of the initial directive from the top. As a consequence, greater controls have to be instituted in order to prevent such possible garbling of communications along the hierarchical chain.

---

[9] By this we mean the number of steps in a production process. If we increase specialization we increase the number of men involved in the basic work group, and in this sense the number of steps through which, say, the raw material has to pass before it becomes a commodity.

All in all, our analysis suggests (and it is only intended to be suggestive) that with greater specialization there is likely to be an increase in the complexity of the organization and a more than proportional increase in the control costs necessary to avoid lapses from some standard of expected performance.

Another way of looking at the matter is to see the points at which control is applied, and to observe how the control points increase with increases in specialization. As specialization grows, so does the number of junctures that exist with activity bundles. This means that the number of interdependencies between different individuals is increased, as well as the number of possibilities of external control at each juncture.

A similar argument holds for roles other than the performance roles. As an organization becomes more complex the number of individuals in, say, the direction complex, or the player-determination complex, will increase and the potential connections between the player-determining decisions and the direction decisions will increase at least proportionately. Here, clearly, we have additional possibilities for control, and hence additional possible control costs. Throughout these considerations it is important to keep in mind that on the one hand control opportunities and control activities are likely to increase with the degree of specialization, but on the other hand the opportunities for economies of scale *within* the control complex itself are simultaneously increased. It is an empirical question whether these growing opportunities of economies of scale are more or less significant than the growth in control activities.

CHANGING ORGANIZATION STRUCTURES, SPECIALIZATION, AND SCALE

The previous discussion was carried out on the basis of a constant span of responsibility. But the span of responsibility is a variable, or at least so we shall argue. We consider therefore the possibility that it will often pay to change the span (or spans) of responsibility, as well as other aspects of the organization structure, when the degree of specialization and the scale of operation change.

To begin with, we have to distinguish between optimally loaded and nonoptimally loaded spans of responsibility. A span of responsibility is optimally loaded if an increase in the span increases inefficiency to a greater extent than it decreases costs; i.e., the desirability of a given span of responsibility depends on two considerations: its costs and its efficiency. (1) As the ratio of subordinates to superior increases, the cost of responsibility (or direction, etc.) per subordinate declines. (2) But as the subordinate-superior ratio increases (at least beyond some point), the efficiency of supervision declines.[10] Direction or supervision at some point impinges on output. The efficiency of supervision is a derived efficiency—derived from its effect on ultimate performance. The important thing is not how closely a man is supervised, commanded, directed, or controlled, but rather how the supervision, etc., affects output. We shall employ the concept of the efficiency of the span of responsibility in this sense. Balancing against each other the efficiency and cost elements we can determine the relative desirabilities of different (average) spans of responsibility.[11] From this viewpoint a span of responsibility may be either underloaded, optimally loaded, or overloaded.

Now, the structure of the organization will depend on the spans of responsibility at various hierarchical levels within the organization. We cannot go into all possible relations between spans, structure, and efficiency at this stage. Our main point here is to suggest that the optimum span of responsibility is a variable (not a constant) that depends on the degree of specialization, the scale of operations, and the qualities of personnel available to the firm. On the basis of such considerations we can show that the numbers of levels in the hierarchy need not increase with increases in degree of specialization and scale, nor need costs of administration per unit of output necessarily grow with such changes.

---

[10] Assuming that the superior does not prevent, or diminish, the subordinate's effectiveness on the job.

"Supervision" is one type of responsibility that comes under the rubric of the "span of responsibility." We use it here, mostly for want of a better term, to suggest what we mean by the efficiency of the span of responsibility.

[11] The span of responsibility will normally vary in different sectors of the same firm. However, for purposes of the present discussion we need not take such variations into account. The reader may interpret the pages that follow in terms of the average or representative span of responsibility within the firm.

Consider first the possibility of not increasing the number of hierarchical levels with expansion. For example, suppose that at the outset the spans of responsibility are all underloaded. Here it would be possible, clearly, to add to the scale of operations without increasing administrative staff. But why should the span of responsibility ever have been underloaded with the smaller scale of operations? Entirely apart from ignorance and error, underloading is possible on other grounds. At a small scale of operations it may be more economic to employ higher quality and more specialized supervisors and keep them less than fully loaded; i.e., the higher cost of the better quality supervisors may be more than offset by their greater effectiveness.[12] Of course, we need not go to the extreme of postulating that all administrators are underloaded. As long as some are underloaded it may be possible to increase scale without increasing hierarchy or administrative costs per unit of output.

A somewhat similar argument can be advanced even if we begin with an optimal span of responsibility for all supervisory posts at all hierarchical levels; i.e., with larger scale and a greater degree of specialization it may pay to employ better trained and more specialized administrators whose optimally loaded span of responsibility is greater than that of administrators most economically employed at a smaller scale and lesser degree of specialization. As a result there is no necessary increase in the number of hierarchical levels. For example, at the smaller scale (and associated lesser degree of specialization) the optimal span may be one to four, using administrators who give fairly detailed directions of a varied nature. With twice the scale (and specialization) it pays to hire more specialized, higher quality supervisors, who give less detailed directions, within a narrower range, and are therefore able to administer, say, eight workers on the average rather than four. In addition, with a greater degree of specialization the workers carry out a narrower range of activities and may therefore require less supervision.

---

[12] But why not use lower quality supervisors in the first place and load them up fully? This may not be possible at the smaller scale of operations. Higher quality supervisors may have more specialized capacities, and to load them fully may involve saddling them with such a variety of subordinates that their effectiveness would decline below the point where it would be worth while to have them.

A number of other variations and possibilities on the above theme could be worked out. But we need not labor this point at the present juncture.

REPETITION, DUPLICATION, AND COSTS

The reader will have observed that we have couched much of the previous discussion in terms of likelihoods and possibilities. This emphasis was quite intentional. The reason, as already suggested, is that it is not possible on a priori grounds to prove that costs of production must rise with expansion as long as equally good factors of production remain available to the firm as it expands. Whether or not costs rise with greater scale of operations will depend on the appropriateness of the organization structure at various sizes. But since in some industries, or activity areas, organizations and firms certainly do become less efficient (beyond some point) with increases in size, it was necessary to examine why. At the other extreme, it is important to note, as Kaldor[13] pointed out, that increasing costs can *almost* always be avoided by duplication. That is to say, to avoid rising costs with increases in scale we can always attempt to duplicate optimally organized plants. The possibility and desirability of duplication need not be applied only to plants. It could be applied to almost any optimal substructure within an organization.

The notions of repetition and duplication are exceedingly important to problems of organization. This is a theme that we have, at the very least, considered implicitly in previous sections, and one that we shall refer to again as we proceed. Duplication and repetition are similar phenomena operating in different dimensions. Repetition is duplication in the time dimension, or duplication may be looked at as repetition in the space domain.

---

[13] N. Kaldor, "The Equilibrium of the Firm," *Economic Journal*, March, 1934, p. 60.

Now, the virtue of repetition and duplication is that one can use to advantage previously successful experiences. In a world of uncertainty it is useful to know that once one has discovered something that works pretty well, one can use it again and again. This is one way of cashing in, so to speak, on the investment in exploratory and trial-and-error behavior, or the costs of experience.[14] Similarly, if a highly satisfactory organizational structure is discovered, then rather than incur further exploratory costs of discovering a new structure for a larger-scale organization it may pay simply to duplicate the existing organization. Duplication reduces some of the uncertainty costs of new organizational forms. Instead of having one organization of scale X we can have two or more substructures each of scale X, each one duplicating the structure of the initial organization. Thus increasing costs can be avoided to a considerable extent. Of course, before duplicating an existing structure we would want to know whether it is optimal so that there are no obviously superior structures that could lead to economies of scale. In others words, the existing smaller structure would have to have exhausted the economies of further specialization and scale. What we would be duplicating then would be parallel substructures.

But would not such substructures have to be coördinated, and would we not then come up against increasing costs of coördination at the top? There are three possibilities: (1) The factors necessary for the coördination of the duplicated structures may be trivial. (2) Some efficiency advantages may accrue to the *set* of duplicated substructures. (3) There may always be a combined structure consisting of a constellation of duplicated substructures that is more efficient than a smaller combined structure.

Possibilities (1) and (2) may be considered simultaneously. Suppose that we duplicate plants of optimal size and in a sense optimally organized. As plants are added there has to be some top administration above the plant level. But the cost of additional top-level administrators may be so small in proportion to the output of an additional plant that the increase in cost per unit of output is trivial. In addition, the advantages of specialized services at or near top administrative levels (e.g., with respect to finance) may be sufficient to counteract the greater administrative cost per unit that would have occurred otherwise.

---

[14] See Chapter 19 for the use of the notion.

In this context it is of interest to consider what might be called the ultimate limited factor thesis. This is the view that has been advanced by Kaldor,[15] among others. Broadly speaking, it holds that there is only "a given unit of coördinating ability," at the top. Even though, say, the board of directors of a corporation can be *expanded,* it cannot do any more decision-making than a "single brain." In other words, it is this single-brain limitation that forms the fixed factor of the firm in the long run. But is there an ultimate coördinating or decision-making capacity or a "single brain" limitation? Let us look at this from the point of view of some of the ideas advanced in previous chapters.

First, no one in an organization necessarily makes ultimate *detailed* decisions. Nor do decisions at the top have to be made in any *specific* degree of detail. In other words, there is no fixed span of command or responsibility, etc., at the top. As previously argued, the optimal span of command (or supervision, etc.) is not a constant, but a variable. In the chapter on authority and communication we saw that by widening the alternative segments in a top executive's field of action we could increase his "span of decision-making." That is, by making more general decisions he could make decisions on a wider front, so to speak, affecting more and more people down the line. A top decision-maker is, in a sense, infinitely expansible. He can make decisions involving a greater range of phenomena by making decisions that are more general and that need less detailed information. His powers can be augmented by increasing the staff that analyzes and sifts information for him. In other words, a greater decision-making range at the top can be obtained

---

[15] *Op. cit.* This view is also presented by G. Stigler in *The Theory of Competitive Price,* The Macmillan Company, New York, 1942, p. 138. Interestingly, it is not found in the revised version, *The Theory of Price.*

by widening the alternative segments and thereby requiring less detailed information for decisions. At the same time this often requires a greater degree of control of subordinates' activities down the line. Such an expanded span at the top may be more or less efficient from the point of view of the organization as a whole than a smaller span.[16] The only point we are trying to make here is that there is no a priori reason to believe that this span is in any respect fixed or limited or is the ultimate limiting factor to the expansion of the firm.

---

[16] The decision-making mechanism will usually be larger and contain more specialized parts in the larger organization. We have already argued that there are both advantages and disadvantages to such arrangements, and we cannot determine on purely a priori grounds that the disadvantages must have the edge over the advantages beyond some scale of operations.

15

Considerations Bearing on the Determination of an Efficient Organization Structure

---

SUPPOSE that we were suddenly made the temporary dictator of an organization, so that we could organize or reorganize it in any way we pleased, beginning from scratch so to speak. What sort of principles would we seek to follow and what considerations would we have to keep in mind? The problem could be divided into three aspects: First, we would seek to work out a role structure that, as far as we were able to determine, would maximize performance. Second, we would have to qualify the role structure in order to meet certain necessary conditions or considerations of a purely organizational nature (rather than conditions related directly to performance). Third, we would have to temper the organizational structure in order to meet certain conditions with respect to anticipated (though not completely certain) types of changes in the external environment, or in order to take into account the possibility of the growth or decline of the organization. We consider each of these separately.

PERFORMANCE CRITERIA

Our task here is to arrange the role structure in such a way as to maximize performance. At the outset we shall assume that each role has only one interpretation. In this case the question boils down to the determination of the fields of action that should belong to each role. Later in our argument we shall drop this assumption and consider the possibility of multiple interpretations.

Under our simplified assumption the optimal conditions for the role structure can be stated very easily in a parallel manner to that used in determining the optimal allocation of resources.[1] The problem breaks down into two parts: (1) Under what conditions should you add a field of action to the organization or delete it therefrom? And (2) under what conditions should you shift a field of action from one role to another? Now the primary consideration has to do with the assessment of how a field of action in a given role will affect performance. In other words, a rational decision must depend on a judgment, calculation, or guess as to how a field of action within some given role will affect performance. Two elements are involved. On the one hand, by adding a field of action to a given role we may increase the performance of the organization in some way. For example, adding a certain field of action to a given role may be expected to increase the gross revenue of the firm in some specified manner. On the other hand, adding a field of action may entail a cost. It will usually be necessary to pay a player in a role with an additional field of action a higher amount than if the role did not possess the field of action under consideration. Thus, we balance our assessment of the additional costs of an additional field of action against the expected gain from it.

The following are procedures by which we might obtain an efficient role structure: (1) We should add a field of action if, in at least one role, it adds more to output or performance than it costs. Similarly, we should delete a field of action from the role structure if, in *every* role, it costs more than it produces. (2) We should shift the field of action from role A to role B if the net addition to productivity is greater when the field of action is in role B rather than in role A. We would keep shifting fields of action from one role to another until no more gains could be obtained by any more shifts—that is, until every field of action produced in its role the same as or more than it could produce in any other role.

---

[1] See Chapters 2 to 5 for the logic of determining optimal allocations and some of its applications.

However, we have argued in previous chapters that in most cases it is unwise to have roles with only one interpretation. There are advantages to permitting the player to determine the nature of the role to some extent. Thus, the question of determining how many interpretations should be permitted in a role is also a matter of balancing the advantages of an additional interpretation against the disadvantages thereof.

To look into this matter of role interpretations, let us examine the disadvantages or costs of having single or unique role interpretations:

1. With only a single role interpretation, we ought to be fairly certain that the role has been established on the basis of the maximum amount of information possible with respect to its productivity, its ability to mesh easily with other roles, and its flexibility in time of change. But the maximum amount of information may not be available—indeed, it usually is not available—to those determining the role structure. It is probably intrinsic to the nature of the problem that additional useful information can be obtained only by actually playing the role. This means, of course, that we must allow for changes within the role as a result of experience derived from playing it. Hence, it would be wiser at the outset to permit more than one interpretation so that this additional information obtained by a role-player could be utilized.

2. Not all players within a role will be the exact duplicates of each other. This suggests the desirability of having more than one

permissible interpretation that might be chosen by a player to suit his particular characteristics and capacities.[2]

3. With unique role interpretations we would have to be very sure that the roles were arranged so as to mesh perfectly with each other. On the other hand, with more than one interpretation players could use the additional information obtained in playing the role to choose an interpretation that would mesh with other roles. This, of course, is part of the advantage considered under (1) above.

4. Unique role interpretations are likely to be very inflexible with respect to those changes in the environment that would make it desirable to have changes in roles. With multiple role interpretations some of these changes can be handled informally simply by having the players shift from one permissible role interpretation to another permissible role interpretation.

5. A unique role interpretation does not leave room for exploratory behavior on the part of the player, whereas permitting exploratory behavior may enable the player to choose a superior role interpretation so that the productivity of the organization is increased.

The main point to be considered here is that, given the limited information available, those in a position to determine the role structure would be unable to choose between the various role interpretations. In other words, they would be forced to be indifferent with respect to a number of role interpretations. Consequently it would seem wise, up to a point, to leave these interpretations open so that the player might be given a chance to exercise some discretion in the matter. Of course, if the player is to choose wisely, incentives must be arranged in such a way that he is induced to choose the best role interpretation within his power. As much as possible, then, the incentives should be positively correlated with performance in terms of the appropriate objectives of the organization.[3]

---

[2] In practice this problem is solved in part by careful player selection and by psychological and other types of testing to see that a player's characteristics are consistent with a rather narrow range of role interpretations.

[3] Cf. Chapter 12 on motivation on this point.

At the same time it is not true that the greater the number of possible interpretations the better. Clearly, as we increase the number of interpretations we are likely to incur some of the disadvantages of having too many interpretations. (1) If we have too many interpretations, an additional interpretation will increase the chance of a player's choosing a nonoptimal interpretation. (2) Beyond some point an additional interpretation will increase the chance that the players' voluntarily chosen interpretations will not mesh. In other words, suppose that every role has a large number of interretations. The players appointed will choose, as a group, one interpretation from each role. Some of these sets of interpretations will mesh, but it is possible that many of the combinations will not. The greater the number of combinations, the greater the chance of finding a combination that will not mesh and the more costly it might be to correct the nonmeshing situation through trial and error or other means. (3) Beyond some point the greater the number of interpretations the greater the difficulty of finding people who have the powers of discretion and flexibility to settle on an optimal role interpretation. (4) Where specific choice rules are to be set up for various players in their roles, it seems likely that beyond a certain point the greater the number of interpretations to a role, the more difficult it would be to determine and communicate the necessary rules of choice to be used within the role. This is not intended to be an exhaustive list, but only to indicate the nature of the considerations that may be involved.

We see that there are both advantages and disadvantages to adding an additional interpretation to a role. To approach an optimum, an interpretation should be added to any given role as long as the additional advantage is greater than the additional disadvantage. But certain considerations must be kept in mind. We must consider not only the net advantage of an additional interpretation to the performance of the role in question, but also how it will affect the productivity of the other roles that are related to this role, and to what extent it may require a greater or lesser number of fields of action within other roles in order to support the interpretation within this role.

In actual practice it is unlikely that a role and all of its interpretations will be spelled out in detail to a player. On the contrary, the most important duties and responsibilities will be stated with some precision, other expectations of performance will be set out with varying degrees of vagueness, and finally, certain prohibitions may be declared or implied to indicate the bounds beyond which the role-player must not go. Such a communication will give the average player enough scope to get started in the role, and gradually, through a process of trial and error, as well as interaction with other members of the organization, he will settle on what seems to him the best role interpretation.

This summary suggests the considerations to be kept in mind when the nature of a role is communicated to an individual. The use of vagueness and prohibitions will imply a fairly wide number of implicit role interpretations, whereas a very detailed job description will probably indicate a narrow range of role interpretations. The point is not so much that specificity versus vagueness determines the range of interpretations, but rather that in describing a role we must realize how vagueness, specificity, permissive statements, and prohibitions may be used to suggest the range of permissible interpretations. In other words, we should not imagine that a vague description of duties and responsibilities allows us to expect performance in accordance with a unique role interpretation.

ORGANIZATIONAL CRITERIA

At the same time that the performance criteria of an optimal role structure are applied, we should keep in mind that certain organizational criteria must also be met. If they are not met, then either performance will not be optimal, or the optimal structure will not be stable, or the organizational structure will not be sufficiently stable with respect to changes. A discussion of some of the main organizational criteria to be considered follows:

1. The appropriate objective for the organization should also be the objective of the direction role complex and should not be in conflict with any other objectives of those in the direction role com-

plex. One question that immediately arises is, What is the appropriate objective of an organization? Our aim in this section is not really to answer this question. Rather, the approach to be taken is that if there is such a thing in any specific case as an appropriate objective of an organization, this should also be the objective of the direction role complex. In a great many cases it is clear that we could quite accurately speak of the correct objective of the organization. How this is to be determined is not the main problem of the moment, although often enough there is evidence to suggest a solution. For example, the ones who are to determine the objectives of the organization may be established by law—there will very often be a governing body such as a board of directors, or a board of trustees, or a court—or set forth in a charter granted to the organization or in the constitution of the organization, and so on. The point is that the legitimate objectives within the existing societal arrangements can often be clearly determined. However, in some organizations the objectives are far from clear. In such cases we will not be able to speak of the efficient performance of the organization unless we can assume a certain objective, or unless we qualify our assessment of performance by an assumption of some objective while indicating that this assumption may not hold. In any event, in the following discussion we shall simply assume that there is such a thing as an appropriate objective or set of objectives and shall examine our organizational criteria from that point of view.

Broadly put, there are three general organizational criteria that should be met: (1) The power structure of the organization should be such that the objectives of those in power are consistent with the objectives of the organization; where they are inconsistent, controls should be such that organizational objectives rather than personal objectives are generally followed. (2) The proper organizational power structure[4] should be stable. (3) The role structure and power structure should be flexible enough to adjust to changes in the environment.

---

[4] More than one power structure may be proper in the sense that it meets criterion (1) above.

We now consider our first organizational criterion in detail. The power structure of the organization will usually depend on the nature and performance of two role complexes—the role-determining role complex and the player-determining role complex. We know intuitively that the role-determining complex is of primary importance from the point of view of influencing the power relationships within the organization. The reason stems directly from the concept of power itself, which is essentially a relationship between two or more roles. This relationship will naturally be unidirectional (asymmetric) and will involve the concept of authority. Such authoritative relations are usually defined with respect to only a range of situations and not with respect to all situations. Only in slave societies are they established with respect to all possible situations. Now, since authoritative situations are relations between roles, and since power depends on a complex of authoritative situations, and since the role-determining complex determines the nature of the roles involved, it follows the role-determining complex can be exceedingly significant in determining the power structure of the organization but it is especially significant in determining the power structures that can arise within the organization. The objectives of the people in the role-determining complex, then, must not be inconsistent with those of the group who determine the objectives of the organization.

Let us look at this matter from the point of view of a private corporation. The stockholders may be said to be the group that determines the objectives of the organization. They presumably desire the maximization of the profits of the organization. However, stockholders do not act for themselves, but only through a board of directors. It therefore follows that the board of directors must be chosen in such a way that they can carry out the objectives of the stockholders and that they are subject to sanctions and controls from the stockholders if they do not do so. They in turn must arrange that the directing officers of the corporation also follow the objectives of the organization, and so on down the line of the hierarchy. If in some way an intermediate group can follow other objectives without controls or sanctions, then the power structure may be said to be one that is inconsistent with organizational aims.[5] Whether or not this is the case will generally depend on the power that exists within a role (and hence indirectly on the role-determining complex that determined the role) or within a group of roles that can form a coalition, and also on the extent and nature of the controls that exist to prevent those in various roles from following nonorganizational objectives.

Since in the usual case the nature of a role is to some extent determined by players, the player-selection roles are also of crucial significance in determining the organizational power structure. That is to say, the players chosen should not be those whose interests are likely to differ from those of the organization. A recognition of this postulate is usually shown by governments, where an individual cannot be appointed to a certain post if his known private interests differ from the interests he is supposed to pursue within that post.

We now consider some subsidiary points along the same lines. The organizational objectives should be interpreted by those in the direction role complex rather than in other role groups. It is the directional role complex that is responsible for performance, and if another group determines organizational objectives, then there may be a power conflict between it and the directional role complex. In other words, performance will be different if the directional role complex does not follow the organizational objectives.

Another consideration is that control must be subsidiary to direction. Otherwise, those in the control role may enforce directions and performances in opposition to decisions made by those in the directional role complex. At the same time, those in the directional complex must be able to communicate their directives, or objectives, or rules of selection, or some combination of these. Thus the necessary linkage roles must also be subsidiary to the directional role complex. But the communication of directives is not sufficient. Those in directional roles must have recognized authority so that when necessary they can control and impose the desired standards of performance; hence our initial assertion that control activities should be subsidiary to the interests of directional activities.

---

[5] Assuming, as we have posited previously in this chapter, that the organizational aims are given at the outset, and that we do not always equate legitimate organizational aims with the aims of the group that holds power.

Controls are also related to the payoff function since the payoff is one way of arranging bonuses and sanctions in order to stimulate desired performance, deter undesirable performance, and in part help establish desirable motivations within the organization. The control mechanism, therefore, must in some way be considered in determining the payoff scheme. However, the payoff scheme should in no way remain subsidiary to the control mechanism.

It may often be economical to use non-payoff substitutes as part of the control mechanism where these are less expensive than using the payoff scheme.[6] Such substitutes may involve the use of propaganda or pep talks to boost morale. In many cases medals, titles, or reprimands can be used with effect in this connection. For example, it sometimes happens in universities without a fixed salary structure that a staff member is given an increase in rank instead of an increase in salary. Such schemes, of course, can be used on a generalized as well as on a specific level. In many countries honors and titles are bestowed on people in recognition of performance, although they may not carry with them any material perquisite. However, where honors and titles lack material perquisites they may not be especially effective; they could not be used to stimulate the efforts of individuals who do not care for such honors, but who nevertheless ought to be stimulated from the point of view of organizational objectives. As a consequence it may be difficult to use a generalized scheme of substituting honors for material payoffs, since there may be groups who will not respond to this scheme. Some compromise between the two approaches will usually prove to be optimal.

---

[6] Expense is not the only consideration. In some cases it may not be practical to give monetary rewards for some desirable types of performance; e.g., there may be no known way to assess the value of an improved role interpretation, but we may want to encourage the experimentation with and discovery of improved role interpretations. Since monetary rewards cannot be rationalized in many such cases, devices to show approval, such as commendations of one sort or another, may be more sensible than attempts to bestow monetary rewards.

The foregoing paragraphs should not be taken to imply that all types of roles should be subsidiary to the direction role. Indeed, this conclusion can easily be shown to be faulty from an organizational point of view. In brief, such a procedure would make those in the direction roles too powerful and enable them to pursue objectives other than organizational objectives. It is especially important that some degree of autonomy be given to those in the payoff-determining roles and to those in the player-selection roles. The main notion here is that it would be dangerous from the point of view of the organizational objectives for both payoff determination and player determination to fall simultaneously into the "wrong" hands.

Let us consider for a moment the possibility of having payoff determination and player determination fall into the "wrong" hands in order to see what is involved. Suppose both of these roles are in the hands of those in the directional complex. Suppose further that a hierarchical scheme exists so that real power resides in the hands of a small group at the apex of the hierarchical pyramid. Now, those in the direction role complex will normally receive their payoff on the basis of their personal performance. But if they control the payoff-determination roles, then they also control the payoff-determination scheme, and as a consequence they can reward themselves entirely apart from performance. Since maximum performance may involve efforts on their parts that they find burdensome, they now have a means of achieving the same rewards without assuming the burdens. In addition, if they control the player-determination roles, they also can be sure of remaining in their favored position. In other words, they are in a position to milk the organization for their own benefit at the expense of organizational aims and/or at the expense of others within the organization.

Now let us see how the power of such a group is weakened if we separate certain role complexes from their control. First, suppose that we give a very high degree of autonomy to the player-determination roles. Even though the direction role group controls the payoff scheme, it is no longer assured of being in a position to gain from the payoff scheme it works out. For example, even though the members of the group have arranged for certain roles to receive excessive rewards, they can no longer be sure they will remain in those roles. Second, if the payoff role complex is separated from the directional roles and given a very high degree of autonomy, then we can see that the initial advantages possessed by those in the directional roles will no longer exist.

The reader will observe that we use the words "a high degree of autonomy" rather than "complete autonomy." The reason is that too high a price can be paid by having overelaborate controls to assure the pursuit of organizational objectives. To be specific, complete autonomy for the direction roles, the payoff-determination roles, and the player-determination roles would be unlikely to maximize the value of performance. We have already argued in previous sections that payoffs should be set so as to stimulate performance. Some payoff arrangements affect performance more favorably than others, and as a consequence those who direct performance should also have a hand in determining the payoff scheme so as to coördinate the scheme with performance. The main notion here is that those in direction roles are likely to accumulate information and insights into the relation between payoffs and performance that would not be immediately available, or available to the same degree, to an autonomous group of payoff determiners. The same will hold even more forcibly with respect to player determination. Those in direct charge of performance are best able to tell whether or not a player is handling his task adequately. Thus, some degree of coördination between these three role complexes is absolutely necessary. On the other hand, it is also desirable that the three role complexes be sufficiently autonomous to reduce the incentives for those in the various roles not to follow the organizational objectives.

A similar analysis can be carried out between any of the roles we have just considered and the role-determination complex. Control of role determination and player determination would be enough for any group to arrange matters to their own advantage rather than to the organization's advantage. A group that can manipulate the role structure and put themselves in the most desired roles can clearly do almost anything they want to with an organization. The reason is that control of the role structure implies an ability to control payoff determination, since roles can be so rearranged that payoff determination and role determination are in the same hands. Clearly, role determination must be autonomous from player determination. However, both have to be coördinated with the direction role complex, since it is the latter that is in a position to determine how effective the role structure really is. In sum, one might say that, while all of the role complexes must be coördinated through direction so that their effectiveness can be judged in terms of performance on the basis of the organization's objectives, there must be sufficient autonomy between role complexes so that the pursuit of nonorganizational objectives is made both more difficult and less profitable than otherwise. As usual, considerations on both sides of the picture have to be balanced against each other.

It may perhaps be argued that it is not necessary for roles to be autonomous as long as no individual player handles such a collection of fields of action within his role that he is powerful enough not to follow the organizational objectives. In other words, as long as the roles are shared among individuals, even though each individual carries out, in some respects, fields of action from all of the role types that we have considered, there is no danger of a thwarting of the organizational objectives. The extent to which this position is correct, however, will depend on the distribution of fields of action between individuals and the possibility of agreement between individuals to pursue personal objectives at the expense of organizational objectives. A significant difficulty with the approach of a personal distribution of power without considering how role complexes are arranged is that in the former case a situation may arise in which each person permits every other person to pursue personal objectives to some extent, since in this way each is assured that he could pursue his own objectives without too great a fear of reprisal by another. To put the matter colloquially, a scheme of mutual "back scratching" could develop if the separation of the role complexes we have dealt with is not effectuated to some degree.

It is important not only that at the outset the role structure should be organizationally desirable but that if for some reason it becomes organizationally undesirable (in the sense that some parts of the organization do not pursue organizational objectives) machinery can be quickly set in motion to correct the situation. Let us consider how such a situation can arise. It may arise as a consequence of at least one of three events: a change in players, a change in role structure, or a coalition of players that is strong enough to enforce the pursuit of nonorganizational objectives.

The machinery to correct such a disequilibrium would consist of several parts. First, the undesirable state must be perceived. Second, it must be analyzed. And third, action must be taken to correct it. This implies that part of the routine information-gathering facilities of the organization must be of a kind that will show up the fact of the pursuit of antiorganizational objectives when it occurs. This in turn means that the information must be such as will enable officers of the organization to distinguish between lapses from adequate performance that are of a routine nature and performance stemming from pursuit of nonorganizational objectives. The routine analysis of performance must show the distinction we have just mentioned and also enable the analyst to recognize, at least tentatively, cases of possible antiorganizational behavior. Finally, all the information-gathering and -analyzing activities are pointless if they do not lead to remedial action when necessary. If the initial cause of the antiorganizational behavior is a change in players, corrective action could be achieved by further player changes and improved player selection in the future.

If the initial cause of the antiorganizational behavior is a change in the role structure, the problem is much more complex. In other words, if the cause has really to do with the fact that the role structure does not induce organizational behavior or creates antiorganizational motivations, then simply changing players cannot correct the situation. Indeed, the situation may not be correctable under such circumstances. The outcome will depend on whether the change in the role structure was such that there is no incentive for anyone in any role to be interested in the organizational objectives. Although some members of the organization may be interested in the organizational objectives, their power may be too diffuse for them to enforce their will on those pursuing nonorganizational objectives. For example, in a corporation the majority of the stockholders may be so numerous and each of them may hold so few shares that any one individual can do nothing about the mishandling of the corporation's affairs. On the other hand, to form a coalition of such small stockholders would take so much effort that it would not pay any individual to attempt it. In other words, the cost of the effort to a small stockholder in arranging for and gathering the proxies from other small stockholders will usually be very much larger than the possible loss he can undergo as a consequence of the mishandling of the corporation's affairs. There are innumerable cases on record of business corporations which are employed for purposes other than those in the interest of the majority of the stockholders (e.g., the McKesson-Robbins case).

The point is that some role structures are conducive to antiorganizational behavior, and therefore possible initial changes in role structure (beginning with a desirable structure) can lead to antiorganizational behavior. The remedy lies in the first instance in trying to arrange matters so that certain types of changes in the role structure are prohibited. For example, there may be a rule prohibiting a senior officer of a corporation from being on its board of directors. Again, there may be a rule that no role-structure change can be put into effect until review procedures have taken place; thus some assurance is provided that the change will not be conducive to antiorganizational behavior. But if such a change should slip through, there should be additional procedures so that a group within the organization can retroactively recognize the change as an invalid one.

With respect to possible coalitions of players to engage in antiorganizational activities, the best situation would be a preventive scheme whereby such coalitions are made as difficult as possible. But before considering this it is important to note that effective safeguards with respect to antiorganizational behavior caused by player changes or role-structure changes would also prevent or inhibit player coalitions for antiorganizational purposes. The main reason is that the purpose of such a coalition would be to change players or role structures to their advantage.

Probably no scheme can prevent all possible coalitions in the interest of antiorganizational behavior. In other words, there will no doubt always be at least one coalition that can lead to the control of a sufficient segment of decisions so that the coalition can benefit by antiorganizational behavior. However, in some organizational arrangements it is more difficult to establish coalitions than in others. We have to consider some of the characteristics of such organizational arrangements.

The essential aspect is the distribution of authority within the organization. If all authority is in the hands of a small group at the top who behave in an organizational manner, then there is probably little difficulty in maintaining this situation. The only risk would lie in a revolt by those who have no authority whatsoever. However, such a scheme is likely to be very costly. Those at the top are not always in the best position to obtain all of the information needed to make all the detailed decisions, and as a consequence some delegation of authority, and of power, is necessary. The people with this delegated authority then have power over a portion of the organization decisions.

The ultimate source of the problem is that those who can be presumed to act in the interests of the organization and who are the proper determiners of organization policy do not always, either themselves or through their delegates and trustees, make the necessary assessments and reviews of performance to be able to determine when action to prevent further antiorganizational behavior should be taken. The following are some of the conditions under which coalitions might be formed in the interest of pursuing antiorganizational behavior: (1) too great a diffusion of power among those who are the ultimate sources of power—for example, a great diffusion of ownership among the stockholders of a large corporation; (2) too great a diffusion of power among the trustees of power—for example, a very large board of directors with no special interest in the corporation; (3) a coalition among the trustees of power or among the majority of the trustees to pursue their own interests; (4) a coalition among the majority of any group at any layer of authority who are delegated some power to pursue their own interests. All of these cases boil down to two possibilities: Either the power exists but is not used because it is too diffuse, or the trustees of this power form a coalition against the ultimate holders of power.

Even where the ultimate source of power is not diffused, antiorganizational behavior is still possible. For example, there may be no assessment and review of trustee performance. Or there may be no scheme for assessing performance, even though a review procedure might be desired by the ultimate holders of organizational power. Finally, there may be no scheme for determining the responsibility for certain activities of the trustees of power, even though the nature of the results could be assessed. If decisions made by the trustees of power, for example, are made by committees whose composition changes with the situation, there may be no way of assigning responsibility for undesirable decisions to specific members of such committees.

Often the problem is intrinsically difficult. Those without any power of decision are likely to be engaged in fairly homogeneous activities in which case performance is quite easily assessed. But those who are trustees of some degree of power are likely to be engaged in quite heterogeneous activities—i.e., their roles contain many contingent fields of action—so that assessment of performance may be difficult. Furthermore, since the sources of information upon which important decisions are to be taken may be highly diffused, and since expert judgment that would have bearing on the decisions may also be diffused among a large number of individuals, it may seem desirable to make such decisions on the basis of a group consensus or majority opinion. In such cases the assessment of individual responsibilities is very difficult indeed. In other words, how can it be determined whether a bad decision is taken because of poor judgment or in the interest of antiorganizational behavior?

If there are no foolproof remedies for the formation of coalitions, or for tacit agreements in the interest of antiorganizational behavior, there may, nevertheless, be certain things that could be done or built into the role structure so as to reduce the probability of coalitions. It must be remembered that the advantage will accrue to the coalition, since its members will have greater control of a larger number of roles than otherwise and therefore they can gain by manipulating these roles. The remedy lies in diffusing the power among roles as widely as possible and in adequate assessment and review procedures of performance of various roles. Since the source of the problem lies in the necessary delegation of power, the remedy must lie in the retention of power by those delegated to review periodically the performance of their trustees. Thus each layer of authority has to develop for its delegates means of assessing performance, means of independent information-gathering, and periodic review of performance. However, one can clearly go too far in such procedures. Every additional means of control is likely to be more costly to an organization, and, as usual, the costs have to be balanced against the possibilities and consequences of antiorganizational behavior.

ORGANIZATIONAL STRUCTURE AND PERSPECTIVE CHANGE

A third consideration in determining the role structure is the possibility of internally induced or externally imposed changes. We may want one role structure if we expect to maintain that structure permanently, but another if we anticipate that the role structure we will desire next year will be different because next year's circumstances under which the organization is to operate will be different. If we can foresee change, then there should be something in the role structure that will make the organization readily adaptable to change. For example, a role structure suitable for a small organization may not be suitable for a large one. Therefore, the initial role structure should be so arranged as to readily permit the transition to the different role structures appropriate for the different stages of the organization's growth.

Considerations of change imply that the initial role structure should be arranged to take into account the following possibilities: (1) discoveries of opportunities for growth consistent with organizational objectives; (2) discoveries of possible improvements in organizational role structure if the existing role structure is found to be nonoptimal; and (3) changes in the role structure made necessary by the fact that the environment in which the organization operates is changing. The constitution of the organization or the rules, mores, and traditions under which it operates must be such as to permit role-structure and other organizational changes; further, means of assessing periodically the effectiveness of different types of role structure for various conditions should be built into the initial role structure. This would imply some degree of readiness to experiment with different types of role structures so that the desirability of changes can be determined.

A role-structure change will often be to the advantage of some members of the organization and to the disadvantage of others. As a consequence, those (or at least some of those) who may possibly lose by role-structure changes must be insured to some extent against such losses so that they do not have the incentive to obstruct the introduction of the changes.

To the extent that growth can be foreseen, roles should be arranged in such a way that when necessary the capacity of a role can be expanded through duplication—that is, through the assignment of more individuals to that role. Clearly, since duplication is the easiest means of expanding a role, the role should be arranged so that its duplication is possible. In addition, for those roles that are not expandable through duplication, the initial roles should be so designed that players are somewhat underemployed in these roles; then the roles can be expanded when spurts of growth occur.

As with many problems, there are both advantages and disadvantages. In this particular case the advantages of flexibility and responsiveness to change must be balanced against the advantages of traditional behavior with its implied economy of routinized, well-understood, and well-established procedures.

THE DISTRIBUTION OF AUTHORITY

Given the relations between choice, authority, communication, and activities, we can readily see that one of the big problems in creating an organization is to determine the distribution of authority within that organization. Should it all be at the top or should it be diffused as much as possible? From what has already been said we have seen that there are great limitations to keeping all authority at the top if the organization is going to be any larger than some very small size. However, the question of the desired distribution of authority must be based on some criterion or other. As we shall see, the distribution of authority sought is one that will carry out the activities of the organization as inexpensively as possible and still be consistent with the simultaneous considerations of the control of potential antiorganizational behavior and the flexibility that we considered in the last section.

The system of authority distribution is part of the role structure of, and the role relations within, the organization. A "rule follower" will usually be able to carry a larger role, but within a narrower range than a "rule maker." The diffusion of authority will determine the types of players needed in various roles. The complexity of the authority structure will depend to a considerable extent on the degree to which the operation of the organization can be reduced to routine. For example, if the organization is a trading corporation, the diffusion of authority will depend on the degree to which the product can be standardized, on the degree of constancy of demand, on the degree of constancy of the supply of factors, and so on. If everything could be foreseen, and if all operations could be performed at a constant pace, then almost no organization would be necessary. Under such circumstances everyone would repeat the same activities period after period, and the decision-making process would be of the simplest nature.

To look at the matter somewhat differently, we can say that the degree to which the authority system need be complex will depend on the degree to which the operations of the organization can be performed under (1) standing orders, (2) standing rules of choice, (3) standing policies, (4) the degree to which operations require the issuance of new and changing policies, (5) new and changing rules of choice, and (6) new and changing orders. Let us examine these notions in detail.

The most economical form of administration is the standing order. Those who carry out the activities have *specific* orders with respect to what to do; hence, no decision-making is necessary. The only administration would be some routine control procedure to see that the performers were indeed performing in accordance with standing orders. A large group of men could be commanded and controlled by one administrator if they performed entirely under standing orders.

Where standing orders are not possible, some degree of choice has to be exercised in carrying out activities. But the choice situations that arise may be of a type that could be routinized. Under such circumstances it may be possible for the top authority to determine some *standing* rules of choice; i.e., the choice process becomes a routinized procedure. However, it will often pay to specialize this procedure. The ones who apply the rules of choice and hand down more detailed orders may be different from those who perform the ultimate activities. In this case a new echelon is introduced into the hierarchical structure. Also, the cost of players who have the capacity to apply rules of choice may be higher than the cost of those who simply apply standing orders.

Where rules of choice cannot be worked out—that is, where situations are too varied for simple rules—operations can perhaps be carried out under standing *policies*. By a policy we have in mind a series of statements which someone faced with a certain area of decision-making situations can translate into rules of choice or the choices themselves.

We now have three administrative groups: at the bottom, those operating under standing orders; next, a group handling those situations which cannot be reduced to standing orders, who will make decisions on the basis of rules of choice (in some cases the choices may result in a change in the standing orders); third, a group who will translate standing policies into rules of choice as the situations arise. Finally, this leaves for the top administrator the review of standing policies and the creation of new policies, which are then handed down through the hierarchical structure.

Now suppose that no standing orders are possible. That is to say, the day-to-day performance activities must be varied to suit new circumstances. In that case, large-scale organization may still be possible by having a large echelon of "rule appliers." If the information flow entering the organization indicates that the tasks to be performed and the choices to be made can be divided into small segments, then funneling the information flow through a large number of rule appliers will allow specific choices to be made and orders to be given to those engaged in the direct performance activities.

Where no choice rules are possible, large-scale organization may still be achieved by having a large number of policy interpreters. Once again, the information flow would be funneled through policy interpreters, who would then make the more specific choices and, finally, give the more specific orders for performance. It is important to observe that there may be many echelons of policy interpreters and that each lower echelon interprets a more specific segment of policy and translates it into a narrower alternative range than the echelon above.

In very special cases abstract policies cannot be formulated in such a way as to be translated into narrower policies and finally into action. This situation occurs in some of the learned professions. For example, a surgeon may have to work alone and not as part of a large organization that performs his particular type of surgery. However, even this is not a pure case, so to speak. An organization such as a hospital, in which the surgeon is one among a number of specialists, may perform generalized health services better than could each member of the hospital individually.

In all such discussions it is important to keep in mind that both authority and communication are relations between fields of action. Thus, when we solve the problem of distributing the fields of action among the roles in the organization, we at the same time solve the problem of distributing the degrees of authority within the organization. The previous sections have dealt with achieving an optimum distribution of the fields of action. In discussing authority distribution, we discuss a different side of role distribution and not a new element of the situation.

CENTRALIZATION VERSUS DECENTRALIZATION

One of the "big questions" often discussed in textbooks on administration is that of centralization versus decentralization. The preceding sections suggest that the big problem is really the determination of the optimum role structure for the organization. Once we have solved that, centralization versus decentralization is no longer an overall organizational problem. Indeed, one of the difficulties with discussing centralization versus decentralization lies in the fact that the words themselves are exceedingly vague; if defined specifically, as we shall define them, these terms will be seen to be not intrinsically clear-cut because there is an index number problem involved.

What should we mean by the concepts of centralization and decentralization? First, we are concerned with the centralization or decentralization of authority. The extreme instance of complete centralization of authority can be given a specific meaning. In this case *all* decisions, regardless of degree of detail, are made at the top and only at the top, and everybody else carries out detailed orders. We have already considered some of the disadvantages of complete centralization.

Once we leave the extreme situation of complete centralization the question of meaning becomes obscure. What shall we mean by a greater or lesser degree of decentralization? To use an analogy, complete centralization implies that the man at the top, and only he, gets to push all the decision buttons of the organizational machine. A degree of decentralization would imply that people other than the chief executive have a hand in pushing *some* of the decision buttons. But how shall we compare two situations in which some degree of decentralization is involved? In the first case, let us say, more people at the subdictator level get to push some of the decision buttons, but in the second case the fewer people at the subchief level who push some of the decision buttons are pushing more important buttons. Clearly, an index-number problem is involved, one that is insoluble unless we can compare and assign coefficients of relative importance to the different decision buttons. If we cannot do so, then there exists a large category of cases in which we could not say that one form of organization is more or less centralized than another. We shall argue in the ensuing paragraphs that the important cases are likely to be in this category.

Suppose we begin with a firm that is of a very small size and consider the changes in organization as the firm grows. With growth there will *usually* be both centralizing and decentralizing tendencies *simultaneously*. Each scale of firm has an optimal role structure. Of course, during a transitional period the role structure may not be optimal. For our purposes we can omit the transitional effects (that is, the path from one optimal role structure to another) and consider only the differences between the optimal role structures.

At the beginning, our firm is so small as to be composed of operatives who carry out the performance activities and a boss who gives all the detailed orders and therefore makes all the decisions. First, let us see how expansion must result in some degree of decentralization. As the scale increases, the load of detailed decisions becomes so heavy that the single boss can no longer carry it. What he must do is create partial duplicates of himself and coördinate their activities at a higher level. At some point of expansion there will be so many partial duplicates that the task of coördination takes up all of his time. Further expansion is possible only if duplicate organizations of three echelons each are created: the operatives, the first-echelon supervisors, and now a second echelon that coördinates the first echelon, who in turn are coördinated at the top in a general way by the initial boss. This procedure can go on indefinitely. From this point of view we can say that expansion implies greater and greater decentralization. But we do not yet have the whole picture.

With increases in scale it may pay to centralize some activities because of the advantages of specialization. Some functions may now be centralized, because only at the expanded scale of operation would it pay to bring a specialized expert into the firm; for example, personnel selection may be specialized. New functions may be added to the organization which previously were purchased externally; for example, a package-design department may be set up that could work more closely with the firm's needs than is possible by relying on the market provision of package-design services. Also, at a small scale, control of manpower and of activities may be informally exercised by the first-level supervisors. At a larger scale, informal controls will no longer be adequate, and a centralized formal control system may have to be put into force. Indeed, a scheme of internal jurisprudence to handle grievances and conflicts of jurisdiction may have to be created at a very large scale. Finally, centralized analysis and information coördination may be necessary, whereas at a smaller scale such operations are adequately handled at the local level. In part these are for the purpose of avoiding duplication of effort. For example, at a lower scale each unit may have handled its own inventory controls, but at a larger scale it may be desirable to examine and handle some inventory problems from a central viewpoint. This may also be true with respect to new product research or new techniques of production. In any event, what we have said suggests that with growth in scale some activities will be decentralized while others are simultaneously being centralized. Clearly, it would be difficult to say whether, on balance, the total effect is one of greater decentralization or one of greater centralization.

We may look at the same problem from the point of view of the principle of comparative advantage. Suppose that higher echelons of management are more expensive per unit of time than the lower echelons. As expansion takes place it would not pay for the higher echelons to waste their time on less expensive decisions. The latter should obviously be made by less expensive manpower—even if the higher-echelon executive could make these decisions better in an absolute sense. On the other hand, the advantage of specialization may be equally important. If the specialized units are lumpy factors, then with scale increases it becomes more and more possible to take advantage of the specialized lumpy factors.

All of this is not to suggest that questions about centralization versus decentralization are always silly. But with respect to the total organization this is not a good way of posing the organizational problem. The organizational problem is whether or not the role structure is optimal, and if not, how to proceed toward an optimal role structure.

However, with respect to specific activities or activity areas, the question of centralization versus decentralization may be a real one *as a problem of assessment.* In other words, it may in fact be very difficult to tell in specific instances whether the advantage lies with centralization or decentralization. Should every department of a firm handle its own personnel recruitment activities or should they be handled at some central location for all departments? As a firm grows it may be desirable to reassess, at various scales of operation, whether a specific activity should be carried on in the same manner as it was formerly carried on or whether it should be more or less centralized. In the language of our previous analogy, should the decision with respect to this activity be made by a smaller or greater number of decision button pushers than before, under a central authority or with diffused authority? How such questions are decided depends on the assessment of the benefit of centralization versus decentralization. With centralization may be associated some of the advantages of specialization and uniformity of treatment. On the other hand, decentralization decisions can be made closer to the scene of operations, possibly closer to the source of the data necessary for decision-making, and may allow the possibility of tailoring activities and decisions to suit specific circumstances. It is clear that to make assessments of this sort is a meaningful problem with respect to specific activities. However, we cannot easily apply any conclusion to the organization as a whole, because the operations of the organization as a whole involve a host of activities, some of which are better "centralized" and others better "decentralized" as the scale of operations changes.

CONCLUSIONS

We have stressed some of the basic elements that an organization should consider if it is to achieve an optimal role structure.

1. The role structure is determined by the fields of action in each role, the distribution of the fields of action among roles, and the number of interpretations within each role. In general, a field of action should be added to a role if its value to the organization is greater than its cost. Also, a field of action should be shifted from role A to role B if its net value (value to the organization less cost) in B is greater than it is in A.

2. In many cases it is disadvantageous to have roles with only one interpretation. With a single interpretation there is no room for exploratory behavior on the part of a player, there is no room for a player to employ information gathered in the play of his role in order to choose a superior interpretation, the player cannot alter the role to suit his capacities, and so on. On the other hand, it may also be disadvantageous to have a great many interpretations in a role. The greater the number of interpretations, the greater the chance that a nonoptimal interpretation will be chosen. In general, an additional interpretation should be added to a role if its potential advantages are greater than its potential disadvantages.

3. The role structure should reflect certain organizational considerations; e.g., authority, and the potential power that goes with it, should be distributed widely enough to balance the risk of antiorganizational behavior against the costs of the division of authority.

4. Although, in general, most roles should be subservient to the direction roles in order to facilitate the smooth functioning of organizational activities, not *all* of the payoff-determining, player determining, and role-determining roles should be subservient to direction. On the one hand, the complete autonomy of these roles would make direction cumbersome; on the other hand, complete subservience to direction would permit an unnecessary amount of power to be accumulated in the direction roles.

5. If the power structure is a desirable one from the organizational viewpoint, its stability is of significance to the organization. Hence, the organizational structure should have safeguards against changes in the role structure that would more easily permit or be conducive to antiorganizational behavior. Organizational officers must be able to recognize and distinguish between lapses from

optimal (or desired) performance that are due to the poor playing of a role and unsatisfactory performance that is attributable to those antiorganizational pursuits made possible by a change in the role structure.

6. In a similar vein the possibility of coalitions formed in the interest of pursuing antiorganizational objectives should be considered in framing the initial role structure and in determining the limits to be set to future changes in the role structure. Such coalitions may arise because of (a) too great a delegation of authority and power by the ultimate sources of power (and this is often facilitated by a high degree of diffusion of power among the ultimate holders of power—e.g., stockholders) and/or (b) a lack of review of the performance of the trustees of authority and power by those who delegated it. The remedy lies in some degree of diffusion of delegated authority and in the periodic assessment and review of the performance of those with delegated authority. But, as usual, there are costs and disadvantages to an extensive diffusion of delegated authority and extensive review procedures, and these must be balanced against the gain from decreasing the chances of undesirable coalitions.

7. The initial role structure should enable the organization to shift with some degree of ease from one role structure to another in response to potential changes in the role, scope, and scale of the organization. Thus the means of assessing the adequacy of the role structure for organizational purposes, and the ability to change it, should be built into the initial role structure. There are often advantages to be gained from a role structure formed in such a way that the expansion of the organization could be achieved by the duplication of most of the existing roles or by the duplication of many of the existing substructures.

8. The role structure determines the authority distribution within the organization. The authority distribution will determine, in part, the types of players needed by the organization. In addition, the complexity of the authority structure will depend on the extent to which the activities of the organization can be carried out on the basis of routine procedures.

9. The complexity of the authority structure required will depend on the extent to which activities can be carried out on the basis of the following: (a) standing orders, (b) standing rules of choice, (c) standing policies, (d) constantly varying orders, (e) varying rules of choice, or (f) varying policies. As we move from (a) to (f) a greater amount of delegated authority and discretion is required. Usually, the most economical form of operation is likely to be that carried out under standing orders.

10. The problem of centralization versus decentralization is a problem of assessment, not one of organizational structure. The organization problem is to find the optimal role structure. As an organization moves from one scale of operations to another, it is likely simultaneously to decentralize some activities and centralize others. While it may be meaningful to speak of the centralization or the decentralization of a specific activity, it is usually not possible, without getting into a very difficult index-number problem, to speak meaningfully of centralizing or decentralizing the organization as a whole.

Part IV

Implications for Economic Theory

16

The Firm Growth Hypothesis

In the remaining chapters a picture of the firm will be developed around the underlying notions that the factors of production are lumpy and that production takes place by organizing lumpy factors into various constellations under conditions of incomplete knowledge. This approach substitutes for the conventional assumptions of perfect divisibility and perfect knowledge the postulates of lumpiness and incomplete knowledge. That is to say, we shall attempt to derive some of the implications for administrative costs and firm size out of the fact that human beings come in relatively fixed sizes and fixed capacities and that to obtain greater knowledge may involve a cost—either the direct cost of paying for information, or paying for the factors that are employed in producing information, or the indirect cost of experience. These are essentially the costs of communication in the very broad sense. Experience may be looked upon as a means of communicating with nature.

In this chapter we examine why firms often start small and grow in size rather than beginning their operations at an optimum scale and staying there.

INFERENCES FROM THE STATIC THEORY OF THE FIRM

The basis of the theory of the firm under perfectly competitive conditions is the long-run U-shaped average cost curve.[1] It is assumed that up to some scale of operations firms can produce under decreasing costs and beyond that scale only under increasing costs. Therefore, there is some point which is a least-cost point on the curve. It is also assumed that under perfectly competitive conditions all firms have the same technical information and, therefore, are all faced with the same long-run cost curve. We would expect each firm, then, to choose the scale of operations associated with its lowest long-run costs. The inference is that all firms in perfectly competitive industries should be of the same size.[2] But sizes are usually very different within the same industry.

One might argue perhaps that even under perfectly competitive conditions our inference would be true only if all firms were in equilibrium. If disequilibrium conditions are caused by random disturbances to the firms, we might expect such disturbances to have no special bias in their effects, and therefore firms would be more or less evenly distributed around the optimum size. Under such circumstances average firm size would approximate the optimum size. But in fact, the majority of firms are usually smaller than the average size in many industries, and a few firms often dominate the industry.[3]

If all factors, including knowledge, are not equally available on the same terms to all firms, firms can have differing optimum sizes. That is, if firm A possesses superior "managerial talent" to firm B, then A may find that its most efficient plant size (managed by the superior managers) is considerably larger than the most efficient plant size associated with firm B. This explanation of firm size differences is certainly plausible, and it is probably relevant to some actual situations (although it does not answer the point made in the last paragraph). But a shortcoming in this approach lies in the fact that its crucial explanatory element is exogenous to the theory; i.e., it is usually merely assumed that some firms possess unique factors not available to the others. To some extent differential advantages in production held by different firms, such as superior administrative and entrepreneurial setups, arise out of the firms' experience and can be obtained only through experience. This element will be elucidated in what follows.

---

[1] See, for example, George Stigler, *Theory of Price*, The Macmillan Company, New York, 1946, pp. 141 ff.

[2] See Herbert A. Simon and C. P. Bonini, "The Size Distribution of Business Firms," *American Economic Review*, September, 1958, pp. 607–618.

[3] See Chapter 19 for more on this point.

Perhaps of greater importance than the above is the observation, from case histories of various companies, that many of them started as small ventures and grew to be large firms. The existing theory does not really explain why firms should choose to, or have no alternative but to, start small and hope to grow. Our first task, therefore, is to develop a hypothesis consistent with this apparent "preference" for growth rather than the attempt to jump to an optimum scale at the outset.

The general viewpoint behind this chapter is that a firm could pursue a number of strategies when it starts out. The appropriate strategy will, of course, depend on the circumstances in which the firm finds itself. An occasional firm may try to pick the optimum scale at the outset and stay there. However, probably the best strategy for most firms would be to pick a tentative growth path rather than aim at the optimum size immediately.

NATURE OF THE LONG-RUN STRATEGY CHOICE PROBLEM

Let us assume that the firm when starting out seeks a strategy which will maximize the rate of return on capital, and that market and other external conditions remain the same within the time horizon under consideration.[4] The rate of return will depend on the investment stream and on the income stream associated with it. If the firm picks the optimum scale at the outset, then the capital stock will be optimal at the outset, and the income stream produced by the firm, abstracting from seasonal and cyclical variations, will also come in constant amounts each period. On the other hand, if the firm chooses a growth path, the capital stock will be growing and income will also be growing period by period. Should it pick a growing capital size path or a constant capital size path?

---

[4] In other words, we posit that the parameters, both of a market nature and of a technological nature, are assumed by the firm to remain the same period after period. We also abstract from the possibility of a gestation period in the construction of capital.

Under certainty of outcome and under perfect information, we would expect that starting at optimum scale, and therefore investing in all the necessary capital at once, would lead to the greatest rate of profit. Any other strategy would mean that, at least for a time, the firm would be working at less than optimal size and, therefore, earning less than the largest profit rate. Other things equal, we would therefore expect that under certainty assumptions firms would start at the optimum scale and stay there.

Some risk is usually involved in all economic activities. However, there are types of risks that can be reduced considerably by obtaining information, often through exploratory behavior or experience. It is because of these types of risks that a growth path toward the optimum size may often be superior to beginning at a scale that is believed to be the optimum.

At least three types of risks can be reduced by exploration and experience: risks in connection with (1) an incomplete knowledge of the market, (2) an incomplete knowledge of competitors' reactions to the new entrant, and (3) an incomplete knowledge of its organizational capacities by the entering firm itself.

The degree of risk depends on the amount of information available. Up to a point, the greater the amount of information obtained, the less the risk involved. The amount of information in turn depends on what can be obtained through observation and general knowledge and what can be obtained through exploratory behavior and experience. Each time one engages in exploratory behavior one may gain knowledge, but usually not gratis. The costs of exploratory behavior and experience must, therefore, be taken into account in choosing the appropriate long-run strategy.

Whether a firm starts large or small, the initial period will be, in part, exploratory. Part of the losses or less than expected profits occurring in that period may be attributed to the costs of exploration—that is, the costs of obtaining information on how to run the business better. However, a small firm is likely to have smaller exploratory costs than a large firm. The maximum loss involved is greater if the initial capital is large rather than small. Also, if the firm starts out at too large a size, there may be considerable loss involved in the retrenchment and the attempt to become smaller. If it starts at too small a size, there may be little loss involved in becoming larger. In the latter case no assets of the company have to be liquidated at forced-sale prices. From all such considerations it would appear that in most cases the costs of exploration are lower for the small firm than for the larger one. Of course, a firm may possibly start at so small a scale that the type of experience it obtains is not really useful at a later stage.

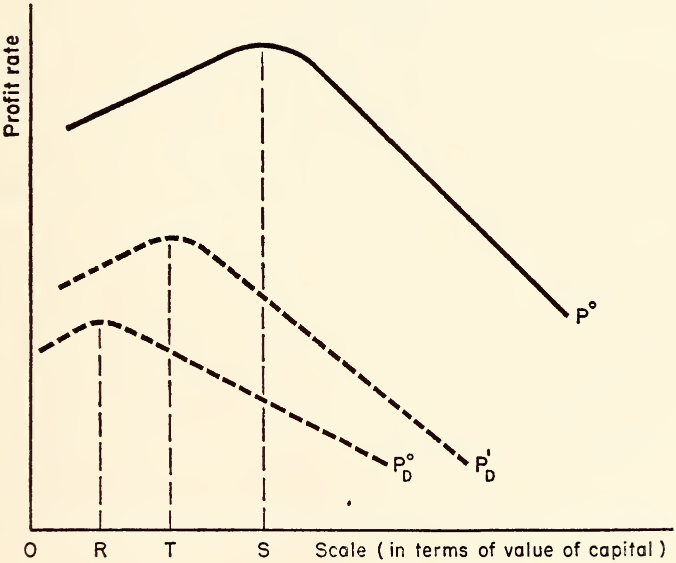


Figure 26.

The general argument can be illustrated with the aid of a diagram. In Figure 26 the curve marked P^0 is the profit curve that associates alternative profit rates with alternative scales of operations. We see that in the early stages it rises and beyond some point it falls off. *OS* is the scale of operations associated with the maximum rate of profit. But this would be true under conditions of maximum knowledge. At the outset it is impossible for the firm to have maximum knowledge. Hence there is a risk involved, and the profit rate has to be discounted accordingly. The broken-line curve $P_D{}^0$ is the discounted profit curve for alternative scales. But we have already argued that the greater the scale the greater the risk and, therefore, the greater the rate of discount. As a consequence the highest discounted profit rate will now be associated with a much smaller scale of operations than under conditions of maximum knowledge. Since at the outset the firm has only a limited knowledge, it would clearly pay for it to start its operations at the lower scale *OR* associated with the highest *discounted* profit rate.

Now, after the firm operates at *OR* and obtains some experience and additional knowledge about market conditions, competitors' reactions, and especially organizational problems and the executives' organizational capacities, it has valuable information which reduces the risk of operations at larger scales of operation. We now depict in Figure 26 a new curve, $P_D{}^1$, which is the discounted profit curve at the end of the first period of operations. This curve is drawn in such a way that in the neighborhood of its current rate of operations there is a smaller gap between the profit rate under maximum knowledge and the current state of knowledge. There is now a new scale at which the firm can earn maximum discounted profits shown by the point *T* in the figure. At this point the firm should grow from *R* to *T*. This procedure can be continued indefinitely until the firm grows to its optimal scale *OS* under maximum knowledge.

The growth path just described may or may not be the best choice for the firm. On the other hand, we must consider the possibility that if the firm had started at *OS* initially it would, of course, have earned lower discounted profits to begin with. But if the anticipated scale of operation is closer to the true optimal scale of operation, then it is possible that the experience gained period by period may be more valuable than that gained via a gradual growth path. Both alternatives are distinct possibilities. What we are primarily interested in showing at this stage is that it is possible for a growth path to lead to a greater discounted profit rate stream than would starting at the anticipated optimal scale. Which situation is the more likely will depend on how ascertainable the optimal scale of operations is without exploratory behavior. In sum, there may be expected to be a great many situations in which the choice of a growth path would be the superior strategy.

COMPETITORS' REACTIONS AND INITIAL SIZE

Considerations of possible reactions by the existing firms in the industry may influence the decision in favor of starting at a scale of operations smaller than what is believed to be the best scale from the purely production point of view.

Suppose that the optimal scale is quite large, so that there is room for relatively few firms within the industry operating at that scale.[5] A new firm wanting to enter this industry might expect that the existing firms would oppose its entry if they thought the new firm would adversely affect their profits. But such opposition would be at a cost to the existing firms. One thing they could do would be to try to drive out the new firm by selling their product at a price below cost until the new entry went bankrupt. This procedure would obviously involve some expense. It would only be worth while to the existing firms if the cut in their anticipated profits caused by the new firm was expected to be greater than the cost to them of a price war. The larger the size of the new firm, the greater the extent to which it would cut into the profits of the existing firms. Therefore, the entering firm might reason that if it started at the optimum size, it would immediately face a price war leading to its own bankruptcy, but that if it started at a very small size, its effect on the other firms' profits would be so small that the larger firms would not engage in a price war. Therefore, the strategy for the new firm might be to start very small and to grow by small amounts every year—each year's increase in size not being sufficient to produce antagonism in the existing firms—eventually reaching its optimum size. To what extent this is a realistic situation might be difficult to say. However, when we add this facet of the situation to those aspects surrounding the relations between information, experience, and risk, we have a total set of circumstances that suggests the advisability of choosing a growth path.

---

[5] Slightly higher "average" costs per unit of output need not be a barrier to entry. There are many considerations involved. See Chapter 19.

THE CAPITAL MARKET, GROWTH, AND SCALE

We may look at the same problem from the point of view of the capital market. Even if an entrepreneur wanted to begin his firm at what he believed to be the optimum scale of operations, he might be unable to get the money to do so. Why, if at the optimal scale the rate of profit is higher than at a smaller scale? Why should he be able to borrow a small but not a very large amount? It must be that the capital market considers the risk much greater at the larger scale of operations than at the smaller. For one thing, it is difficult for an entrepreneur to prove his organizational capacities before the fact, so to speak—before he has organized and operated a large firm. In other words, there are two sides to the coin of experience. Not only does it provide the entrepreneur (or entrepreneurs) with information about his organizational and business capacities and his chances of success, but it provides similar information to the outside world. Quite probably the private assessment of their organizational capacities by most entrepreneurs is somewhat higher than the public assessment. As a consequence, the opportunity open to them is to begin with a small firm using their own capital, or that of close friends, and whatever little they can borrow, and to attempt to grow to optimal size.

Other things equal, larger firms can usually raise more money for growth purposes than smaller ones. The absolute amount of capital that a firm could raise is in part a function of the value of its current assets. The public assessment is likely to operate on the basis of the current performance and the current scale of operations, and perhaps on the basis of the recent rate of growth. But we would not expect the capital market to be willing to invest simply on the basis of a calculation of the anticipated optimal scale of operations which turns out to be very much larger than the current scale. Thus, we find in the operation of the capital market some contributory evidence for our initial hypothesis that, other things equal, the greater the gap between the current scale of operations and the expected optimum scale, the greater the discount of the expected profit rate.

17

Organizational Objectives and the Profit Maximization Hypothesis

---

IN THE last chapter we argued that it may pay for a firm to choose a growth path strategy even where it is interested in long-run profit maximization. But we may raise the question whether large corporate firms do have long-run strategies. Can an organization really be said to have a goal? And if so, how should we interpret statements that characterize the presumed goals of an organization?

One possible viewpoint is that organizations, such as firms, simply behave, but that they cannot be said to have specific goals. There may be some criteria for satisfactory behavior. Also, they may be said to be interested, at the very least, in their survival. But perhaps it cannot really be said that they try to maximize something or other. After all, an organization is a collection of individuals performing within certain roles. It cannot be said to have any behavior apart from the behavior of these individuals. Further, the organization cannot be said to have any aims apart from the aims of the individuals as they play their roles. In other words, an organization does not have a will of its own apart from the wills of the collection of individuals composing it. Therefore, if these individuals do not have a single homogeneous aim, then the organization cannot be said to be pursuing a single aim. If there is nothing about the organizational structure to induce all individuals to have the same aim, the organization itself cannot be said to have a single aim.

The exception to the above would be the case of an organization operating under a dictator with perfect control. If everyone acts only in response to the detailed explicit orders of the dictator, the organization's aims are the same as those of the dictator. For example, a ship may be said to have a certain goal, namely, to arrive at some specific port. But a ship is, in a sense, a dictatorial organization. Supreme power while at sea resides in the captain. However, we have argued that it would be unwise to run most organizations from a purely dictatorial point of view. With corporate firms especially there is a delegation of power from the widely dispersed stockholders to the executives of the corporation. The ship analogy therefore does not apply. We have seen that in any large organization it is desirable that the chief executive, who already is the recipient of delegated powers, delegate authority and discretion to various people below him. There is no reason to expect, then, that in all cases, or even in most cases, choices will be made in terms of the unified aim of the possessors of ultimate power.

Nevertheless, we do often make statements to the effect that firms maximize profits, or that firms aim in a certain direction. Are such statements completely meaningless? Of course, in some cases they are simply incorrect. However, they are not necessarily meaningless. They should not be taken as concrete statements of the definite goals of the entity, the corporation. The corporation as such (in the nondictatorial case) cannot strictly speaking be said to have a goal. We should interpret any such statements as broad descriptions that are implicit in observed organizational behavior. We may argue in specific cases that the firm acts as if it were trying to maximize profits; its behavior, even though we cannot say anything about its goal, is, in general, not inconsistent with profit maximization. The observed behavior may also be not inconsistent with a number of other objectives.

In what follows we shall argue that there is a connection between what we ought reasonably to assume about corporate goals and the nature of the organizational structure. For example, profit maximization may or may not be the most accurate characterization of a corporation's aims. The point of this section is that whether or not this is the case will depend, at least in part, on some characteristics of the organizational structure. In other words, we must not assume, as is often assumed in economic theory, just because an organization is a commercial firm and is presumably organized for private gain, that it necessarily attempts to maximize profits. From this point of view we shall examine, as an example, Professor Baumol's sales maximization hypothesis of oligopoly behavior.[1]

CAREERS VERSUS ORGANIZATIONAL ROLES

Unlike most factors of production, labor cannot always be purchased in the form in which it could be immediately used, nor can it be purchased outright. With the abolition of slavery, human labor can be obtained only in the form of a service and not as a result of purchasing the capital good (i.e., the human being) itself. This unique aspect of the labor factor creates certain organizational problems not involved with other factors. With respect to other factors of production there is an inducement to invest in them even if the return from that investment cannot be obtained immediately. As long as an adequate return can be obtained over the life of the capital good, a sufficient inducement will exist. This, however, is not true with respect to labor, for in most legal systems it is impossible to guarantee to the investor the persistence of the return. To be specific, an investor cannot invest in the education of a worker and insist (*and legally enforce*) that the person work for him under agreed-upon conditions for a definite period. Work contracts are not enforceable. But since the labor is not always available in the form desired, some investment in the selection and training of workers may be unavoidable. To achieve an adequate return from such investment, the organization must devise schemes to induce the workers to remain with it. In other words, arrangements must be made so that the worker is less well off if he leaves the organization that has invested in him than if he stays in it. This, in part, explains the existence of organizational careers.[2]

---

[1] William Baumol, "On the Theory of Oligopoly," *Economica,* August, 1958.

We may define a career as a role sequence with a role goal as an objective. From the point of view of the careerist we visualize a sequence of roles connected with each other in such a way that they form a reasonable path to some desired role—a role that is the ultimate aim of the one traveling through the sequence of roles. For an individual there are often two rewards for successful performance within a given role. First, there is the reward for performance within the role itself (the wage, salary, and/or bonus), and second, there is the reward of promotion to another role along a career path. In examining the problem of the objectives that an organization will actually pursue we must consider whether the career structure induces people to behave in such a way as to be consistent with organizational objectives. In other words, will the inducements of behavior toward role promotion be the same as the inducements of behavior toward performance with respect to the organizational objectives? If they are not the same, will the conflict be resolved in favor of the career objectives or in favor of the organizational objectives?

Another question: Under what conditions will the pursuit of career objectives conflict with presumed organizational objectives? We may expect the career objectives to win out over organizational objectives if some or all of the following conditions hold: (1) if the organizational objective is vague or is not easily determined or interpreted in specific situations; (2) if those who are likely to defend the organizational objectives are widely dispersed and their powers limited and highly diffused—that is, if they could only uphold the organizational objectives by behaving in concert, but not individually; (3) if there is a large division of authority and responsibility within the organization; (4) if assessment of performance is difficult or impossible; (5) if the assessment of successful performance for promotion purposes is easier and different from the assessment of successful organizational performance (e.g., a seniority criterion for promotion). It should be easy to see that if we create a model of an organization with these characteristics it will behave in accordance with the interests of those in it who pursue careers, rather than in accordance with those who loyally pursue only the organizational objectives.

---

[2] See Chapter 12.

A SKETCH OF AN ORGANIZATIONAL MODEL FOR THE SALES MAXIMIZATION HYPOTHESIS

According to Professor Baumol,[3] large firms will attempt to maximize sales revenue subject to a minimum profit constraint rather than to maximize profits. We attempt here to sketch the organizational structure of such large firms so that their resulting implicit behavior will be in conformity with the sales maximization hypothesis.

To set up our model we will start from the top with the initial source of corporate power and work our way to the various lower hierarchical levels. The ultimate source of corporate power and interest lies, of course, with the stockholders. We shall assume that stock ownership is widely diffused. Therefore, the organizational activity that the typical stock owner engages in is to vote for the board of directors. Stockholders cannot usually act in concert without somebody's taking the initiative to collect their proxies. We assume, as is usually the case, that such initiative is quite costly except when carried out by existing management, and as a consequence will be forthcoming only when corporate operations are exceedingly unsatisfactory. We must distinguish here between a satisfactory level of operations and an optimal level. A satisfactory level will be one that will not induce the representatives of the stockholders to change management. The stock owners, of course, act entirely through the directors. These are of two kinds: those who are part of management and have management interests, and those who are outside management. We shall assume that the nonmanagement directors have very wide interests so that their interest in any specific corporation, including the one under consideration, is not a paramount one. Therefore, they too will not make any special efforts to change management as a result of nonoptimal behavior if such efforts are in any way costly to them. We will return to their motivations later.

---

[3] *Op. cit.*, pp. 187–188.

We now consider the management group itself. We assume that they are primarily concerned with their careers as managers. Now, how are they generally to be judged? Will they be judged on the basis of the profit performance of the corporation or on some other basis? It is fairly clear that in a large corporation no one person is likely to be held completely responsible for the progress of the corporation. Usually, there is a considerable division of both responsibility and control. Performance within most roles, other than perhaps those of the president and the chairman of the board, will be assessed in terms of satisfactory performance for some given range of responsibilities. In other words, what will be assessed is usually the thing that is assessable: some index of the size of performance. The quality of performance is usually exceedingly difficult to assess, and as long as performance is satisfactory, the size index will be employed. For example, if a man performs satisfactorily in the role in which 50 men are responsible to him, this performance may be superior to being successful in a role in which only 25 men were responsible to him. In a sense, his career interest is to promote the assessable size of his role. It may be more important to him to perform well in a big role than to perform brilliantly in a small one.[4] Under such circumstances it is clear that he would be interested in as large a role as possible subject to satisfactory performance. Now for most roles, maximizing sales is likely to increase the role or increase the possibility of moving from a smaller role to a bigger one, rather than the alternative of reducing sales and increasing profits. We have thus established the career motivation for sales maximization.

---

[4] This is consistent with the statistical finding that the compensation of executives is more closely related to sales than to profits. See David R. Roberts, "A General Theory of Executive Compensation Based on Statistically Tested Propositions," *Quarterly Journal of Economics*, May 1956, pp. 276 ff. See also H. A. Simon, "The Compensation of Executives," *Sociometry*, March, 1957, pp. 32–35.

Now let us look at the other side of the picture—the control side. Suppose there is some attempt to check on nonoptimal behavior. In the first place, it is easy to determine whether profits are satisfactory—that is, whether they meet the minimal expectations of stockholders. But it is difficult, if not impossible, to determine whether profits could have been higher than they are.[5] In the latter case, it is necessary to be able to compare the actuality with what is usually an untestable possibility. As a consequence there is no way of knowing or it may be very difficult to determine whether the group as a whole has behaved in a nonmaximizing manner.

A further difficulty arises because of possible conflicts between short-run and long-run profit maximization. An act that leads to lower profits today may be deemed desirable in order to achieve larger profits in the future. Sometimes a corporation is reluctant to give up market areas on the grounds that, although they are unprofitable at present, they must be cultivated because of the expectation that they will be profitable in the future. Because the future cannot be foreseen, we cannot tell at the time a decision is made whether such arguments are true or false.

All in all we can see that improvement—that is, increases in sales—may be used as explicit evidence of greater role capacity, not only for subsidiary management personnel, who may achieve a somewhat larger role with sales expansion, but also for top management, since sales expansion is an index of improvement. Furthermore, still further sales expansion is something that could be achieved in the future, whereas indefinite profit expansion may be most difficult to continue. That is to say, the top management may fear very high profits because they may increase stockholders' expectations and, therefore, lead to adverse assessments of management's capacities if the same high profits are not achieved in the future.

---

[5] In some cases a firm may know of actions that it could take that would increase its profits but decrease the scope of its operations. Professor Baumol has pointed out to me cases that have come under his purview as a consultant in which firms refused to give up unprofitable areas of their operations. These are clearly cases where the officers of the firm chose a greater scale of operations in preference to higher profits.

SALES MAXIMIZATION, GROWTH PATHS, AND OPTIMAL SCALE

The foregoing does not solve the question of choosing a growth path versus starting with an optimal size at the outset. One may inquire, if management is interested in maximizing sales, why do they not start with the largest firm size possible consistent with minimum satisfactory profits and thereby achieve maximum sales? It may be desirable to pick a growth path for several reasons. Top management may want improvement period after period so that it can show achievement. If it picked the optimum size to begin with, no future achievement would be possible. But even if the firm wanted to choose the optimum size it would not be able to do so. We have considered the reasons why this is the case in the last chapter and therefore need only mention them briefly. The larger the scale, the larger the uncertainty discount rate, so that for a very large scale the discounted profit rate would drop below the satisfactory level. But in addition, there is the organizational absorption problem to consider. It is difficult to start with a large organization and have it function well without considerable loss during an adjustment period. It would be best, therefore, to start with a small organization and grow period by period so as to absorb the increases in organization most easily. Finally, and perhaps most important, the capital market would not be willing to finance a large-scale organization until management had proved itself at some smaller scale. Even from a career point of view the top managers might not want to risk failure at a large scale when they could feel assured of a reasonable degree of success at a smaller scale and then gradually feel their way toward growth. All of the factors mentioned may not be operating in any specific situation, but enough has been said to suggest that in any typical situation enough of them are likely to operate so that a growth path might readily be preferred to the presumed optimal scale at the outset.

CONCLUSIONS

1. Organizations do not have goals apart from the goals of the individuals that compose them. In fact, at any one time a specific organization need not have any definite goal—except perhaps survival.
2. However, statements about organizational goals are not meaningless. They should be viewed as rough summary statements suggesting how an organization would probably behave when faced with certain choices; i.e., given a choice, the organization should pick that alternative consistent with its presumed goal, if the initial assertion about the organizational goal is correct.
3. The correct organizational goal, using the term in the sense indicated, and organizational behavior consistent with that goal, will depend in part on the role structure of the organization.
4. The formal or legal repositories of ultimate power within an organization, such as stockholders in a corporation, need not be the group that actually wields power, or the group that in fact determines the organizational goal.
5. There need be no uniformity of interest among all the members of an organization. For example, career interests may conflict with presumed organizational interests. It may be in the interest of some members of an organization to behave in a manner inconsistent with the pursuit of presumed organizational aims. Thus a firm may not pursue profit maximization even if it is in the interests of most of the stockholders of that firm to do so.
6. Finally, the actual goals that are pursued by firms cannot be determined on a priori grounds but will depend on the role structure and power structure within the firm—looking at the firm as an organization.

18

Organization Costs and the Scale of Operations

THE FACTS TO BE EXPLAINED

Some general observations and some recent studies[1] suggest that certain trends and broad relationships exist between organizational costs and scale of operations. If our theories of organization are valid, they should, at the very least, lend themselves to explanations that are not inconsistent with such general observations. We therefore consider briefly what some of the empirical findings are, observe that they parallel, to a considerable extent, the empirical findings on the relation between firms, profits, and scale, and try to suggest (in this and the next chapter) how the organizational concepts that we have developed can shed light on these matters.

---

[1] Many of the studies and findings are summarized in Seymour Melman's excellent book, *Dynamic Factors in Industrial Productivity*, Blackwell, Oxford, 1956. Additional findings of this nature are surveyed and summarized in Reinhardt Bendix, *Work and Authority in Industry*, John Wiley and Sons, New York, 1956, chap. 4, "The Bureaucratization of Economic Enterprise," especially pp. 211–226.

The most persistent fact is that in all modern countries there has been a trend toward more and more organization. Specifically, this manifests itself in a higher proportion of administrative personnel to operating personnel, and in a higher ratio of administrative costs to total costs. There can be little doubt, as Seymour Melman's study shows, that both in Great Britain and in the United States the trend continues in an unbroken fashion in most (perhaps all) industries, and for firms of almost all sizes, toward attributing a greater and greater proportion of total costs to administrative costs.

Two additional tendencies are that smaller firms have on the average (1) a higher ratio of administrative costs to total costs and (2) a higher ratio of administrative personnel to operating personnel than larger firms. At the same time, there are often *some* small firms that have as low administrative costs in proportion to total costs as some of the large firms. Finally, it is often observed that during a period of expansion the proportional increase in administrative personnel is significantly greater than the proportional increase in operating personnel.[2]

THE TREND TOWARD GREATER ADMINISTRATIVE COSTS

At first blush it may be thought that the trend toward increasing administrative costs is somehow connected with growing scales of operation. However, the fact that in industry after industry large firms have lower administrative costs in relation to value of output than small firms casts considerable doubt on this hypothesis. For each size class taken separately there is a persistent increasing trend toward higher administration costs. Thus, whatever the explanation, it cannot be entirely in terms of an overall trend toward larger scales of operation.

It has also been suggested that the expansion of government and of government controls, as well as government requests for data, is the source of administrative expansion in commercial firms. Although this may have something to do with the matter, it cannot be a complete explanation in view of the fact that the trend toward greater organization costs seems to have operated as fully in periods when the government was not expanding as in periods when the government was expanding. Furthermore, as a matter of general observation it has not been found that the proportion of administrative costs devoted to meeting government demands for data has been a highly significant proportion of such costs. The trend toward increasing organization costs seems to have persisted just as much in industries that are highly controlled as in those that are not. Thus it would appear that if we are to seek an explanation we must look for it (mostly) elsewhere than in the expansion of government activities.

---

[2] Melman, *op. cit.* See also Samuel E. Hill and Frederick Harbison, *Manpower and Innovation in American Industry*, Princeton University Press, Princeton, 1959, pp. 39–59.

What seems to have occurred over the last half-century, for which figures are available, and probably for a very much longer period for which exact data are difficult to come by, is that there has been a general substitution of machinery for labor and with it a greater need for organizational machinery per direct operative. This suggests the possibility that organization and machinery are complementary factors of production. But why should they be primarily complementary agents of production? Why should we not substitute organization for mechanical machinery just as we have substituted machinery for direct labor? One might argue perhaps that organization itself is a type of machinery and that therefore organizational machinery and mechanical machinery are substitutes one for the other depending on the relative prices of the two. Yet the trend suggests that both organization and nonhuman machinery have grown together and that both have been substituted for direct labor.

We can probably divide what has happened into two effects: a substitution effect and an innovation effect. That is to say, at various times firms have substituted known types of machinery for direct labor when direct labor costs rose compared to machinery costs, and from time to time, as inventions took place, new types of machines were introduced that were also substituted for labor. But why were these mechanical innovations in general of a kind to go together with increasing amounts of organization—why were they not substitutes for organization? To some extent the latter may have been the case, but certainly the predominant effect was not in that direction.

A plausible explanation lies in the likelihood that in most cases methods of production have become more and more specialized, and that the increases in both machinery and organization have been of a kind to facilitate the increasing degree of specialization. The main point is that the trend toward greater labor productivity has been achieved by increasing specialization, and that this in turn has been facilitated by increases in both nonhuman machinery and organizational machinery.[3] Efficiency can often be attributed to repetition and duplication—through processes that enable us to repeat the same activities over and over again. It is this sort of duplication that permits specialization, which in its turn enables us to allocate the factors of production to the performance of those activities in which they are most efficient. In addition, duplication is facilitated by standardization. That is to say, the standardization of a possible variety of activities into a uniform activity creates the conditions necessary for the extension of repetition and duplication. Thus, standardization and duplication are intimately associated with the advantages to be derived from specialization.

Now it is likely that the possibility of the duplication and repetition of relatively simple activities is what stimulates the invention of nonhuman machines to carry out such activities. Thus, the factors entering into the extension of specialization also suggest the type of machinery that can be created to substitute mechanical for human efforts. Usually the simultaneous standardization (and duplication) of the consequences of the standardized activities, namely, the commodities themselves, will be involved. After the standardization of specific activities, the standardization of whole sequences or bundles of activities may also be possible. This too may suggest means of creating nonhuman machinery to carry out some of these sequences of activities. But at the same time we must recall, as we have argued in earlier chapters, that an increase in specialization will imply an increase in the organizational structure. This comes about because specialization permits the separation of the direct performance activities from the direction-control complex. That is to say, specialization enables us to separate out from direct performance those activities that involve the specific direction of the direct activities, the control of such activities, the staff activities associated with the initial determination of the proper set of activities, the linkage activities, and especially those linkage activities necessary for the coördination of the separate direct performance activities. All of this will also involve additional organizational activities because of the larger number of interactions and relations between various sectors of the administrative setup. As a consequence increased specialization is often associated with a more complex organizational structure.

---

[3] And this includes mechanical contrivances to aid the administrative sector. In part there has been a substitution of machinery for administrative labor, but to a much lesser extent than that of machinery for direct labor.

But why should increases in specialization, and the increased efficiency obtained thereby, often necessitate a more complicated organizational structure? Initially we may associate with an increase in specialization an attempt to standardize and duplicate activities as much as possible at the lowest operating level. But this will necessitate coördination at a slightly higher level. However, at the higher coördination level specialization and duplication are also possible, at least to some extent. But the specialization of the first-level administrators implies a second level to coördinate the first level. With increases in scale, higher and higher coördinating levels can be increased to whatever extent necessary depending on the size of the organization and the degree of specialization that has taken place. Clearly, the establishment of a hierarchy is an essential part of the scheme. But this hierarchy may not be, and usually will not be, in the nature of a simple pyramid. In extending the degree of specialization some activities may be decentralized while others are simultaneously centralized. The authority and communication relations need not be simple ones that proceed from top to bottom. They may cut across some layers of administration. In almost all cases the authority relation will be a limited one. Furthermore, some members of the administrative sector may simultaneously carry out some direct performance activities. For example, a foreman may conceivably do some work on a machine for part of the time. Higher officers sometimes perform nonadministrative tasks. In any event, there are many reasons for increased specialization to be associated with increased complexity of organization.

But let us return to our main problem area—the trend toward higher organizational costs. As labor productivity rises and wages rise accordingly, there is a substitution of machinery for labor. At the same time increased output may also be achieved by substituting organization for direct labor. That is to say, increased specialization which yields greater output implies simultaneously an increased amount of organization per direct worker. Probably both the substitution of machinery for labor and the substitution of organization for labor take place simultaneously because there is no reason why a firm should not simultaneously employ both avenues to increase productivity. In addition, we have argued that a great deal of industrial machinery is of a type that facilitates specialization, and since organization both facilitates and is required by specialization, they should be complementary factors of production.

The logic of specialization and its associated elements of standardization and duplication suggests types of machinery that are likely to be efficient. Thus, innovations will be in the direction of fostering a further extension of specialization and a necessary further increase in the degree of organization.

While the foregoing argument is consistent with observed trends, it is not an absolutely necessary argument. It is certainly conceivable that some inventions and innovations should have been of multipurpose machines that decreased rather than increased the degree of specialization. But it is unlikely that such innovations will predominate. After all, it would appear on a priori grounds to be easier to create single-purpose tools than those that will simultaneously perform a variety of activities.

There is one respect in which machinery and administrative personnel may be complementary whatever the nature of the machinery. While the members of the work force who directly operate and tend machines are normally classified as operatives, or direct labor, the firm would, nevertheless, have to employ others to maintain the machines. In other words, the maintenance staff, as well as that portion of the staff concerned with the introduction of machinery and its replacement, would be considered nonoperating personnel. [illegible]

In sum, the explanation of the existing trend is made up of a number of components, although it would be difficult to give a specific weight to each one. These components are the substitution of organization for direct labor, the substitution of machinery for labor, which in turn facilitated greater specialization and hence required greater organization, the "maintenance-complementarity" between machinery and administration, and finally, the likelihood that inventions and innovations were, on balance, of a type that fostered greater rather than less specialization.

FIRM GROWTH AND ADMINISTRATIVE COST

The empirical evidence indicates that large firms will have smaller administrative costs (per unit of output and per operative) on the average than small firms. Yet it is often observed that when a firm expands or plans its expansion the planned and actual increase in administrative personnel is larger proportionately than the increase in output and in operatives. How can we reconcile this paradox? It becomes clear when we see that unlike nonhuman machinery (which can be purchased) the organizational machinery has to be built, and more personnel is usually required for the building of the machine than for its maintenance. In other words, during the expansion period when additional organizational machinery has to be constructed a larger increase in nonoperative personnel is necessary. However, once the expansion is over, the increased administrative personnel has to be either absorbed by the substitution of organization for direct labor (the trend factor discussed in the previous section) or disposed of by the gradual release or lack of replacement of administrative staff upon retirements, dismissals, and resignations. The general trend toward growing administrative staffs is likely to absorb the increases in staff required during the construction period. During such expansionary periods, too, those administrative and machinery innovations will be introduced that imply a higher ratio of administrative to nonadministrative personnel. Both of these factors account for the fact that during the growth spurts the increases in administrative personnel are often larger than the increase in nonadministrative personnel. But since this is likely to hold for the growth spurts of large firms as well as the growth spurts of small firms, it need not affect the average lower administrative/nonadministrative personnel ratios for large firms as against small ones.

FIRM SIZE AND ADMINISTRATIVE COST

Why should large firms have no greater administrative costs per unit of output or per operative than small ones? One would think that increases in scale would bring increased complexity, a more difficult problem of coördination, and greater rather than smaller administrative costs per operative, or per dollar of output. Yet the facts are usually the reverse, at least on the average.

To try to explain this phenomenon we might consider our theme of repetition and duplication and its relation to organization. For very small scales of operation we would expect that the greater the output, the less the administrative cost per unit of output—simply because administrative units come in fixed sizes. Thus, if the firm has only three operatives and the works manager can easily supervise six, it follows that when the firm expands from three to six operatives its administrative cost per unit of output will fall. But suppose that a works manager or foreman can handle no more than six operatives without imperiling efficiency and quality of performance. Clearly, an increase in size of the firm to seven operatives will necessitate an additional foreman and will imply a sharp increase in administrative costs. The answer would appear to be duplication. That is to say, at this level an optimal work unit is one foreman and six operatives. Once the optimal work unit is known, it is obviously best to expand by duplicating optimal work units. Thus the firm should try to expand not from six to seven operatives but in a spurt from six to twelve operatives.

The same ideas can be applied to the next administrative echelon. A manager may be able to handle, let us say, five foremen. A plant with only two work units (of one foreman and six operatives each) would therefore have higher managerial costs than one with five work units. Once the plant size reaches the five-work-units-one-manager level, expansion should, if possible, take place via the duplication of this work complex. Since we have now determined a hierarchy of one manager, five foremen, and thirty operatives as an optimal arrangement at this level, it would seem reasonable for expansion to take place by duplicating such optimal arrangements. For example, a chain store might well expand by opening an additional store of a given optimal size rather than by opening an additional store which was very small and then gradually expanding it.[4]

At the same time, as the firm expands it may find it advantageous to centralize and specialize certain activities, achieving additional savings in administrative costs thereby. On the other hand, a factor that may work in the direction of increasing costs arises out of the possibility of a proportionately larger number of interactions between members of the administrative staff simply because of the larger number of people involved.[5] Thus, although some forces work in the direction of decreasing costs, beyond some point, because of the possibilities of bureaucratic tendencies, some factors may work in the direction of increasing costs.

The relevance of the duplication thesis to questions of administration can, perhaps, be seen from a slightly different point of view:

---

[4] If other organizational schemes associated with still lower costs per unit can be found as output is increased, these, of course, should be and often would be introduced. There is no point in duplication for its own sake. The point is that via duplication we can at least avoid (to a considerable extent) rising administrative costs.

[5] More than twice as many interactions are possible with four people than with two, and so on; i.e., the number of combinations of two people at a time (or of three, etc.) increases much more rapidly than does the number of people. This tendency may be reduced by creating noninteracting or minimally interacting departments.

How can a firm with less than perfect or maximum knowledge use its built-up experience when it has to expand the administrative setup? It must be remembered that beyond a very small size the organizational machine is usually quite a complicated entity. Furthermore, unlike nonhuman machines it cannot be purchased but has to be created on the spot, so to speak. The firm must fall back to some extent on its previous experience in creating such a machine. It will certainly not have a knowledge of the effectiveness of all possible alternative organizational mechanisms. How is it to behave rationally in the face of the partial knowledge and experience in its possession? It would seem that once it has created, through a process of trial and error and perhaps after a number of false starts, a fairly workable and effective "single plant" organization, it should try to expand by duplicating such substructures. From a one-plant firm it should expand, say, into a two-plant firm. In this way it would take advantage of previous experience and probably minimize the cost of creating the additional organization. Again, we may look at an attempt to duplicate successful organizational units as a means of spreading as widely as possible the exploratory costs that must be undergone in the creation of the initial organizational unit.[6]

Now, given some organizational setup that is optimal for a small firm (but not very small), duplication would permit a large firm to do no worse than the small one in terms of administrative costs, if it duplicated the optimal administrative setup by integral multiples to achieve its scale of operations. Very small firms may not do as well as large ones, in part because the indivisibilities in some of the administrative factors would not permit them to create administrative structures that would be as fully utilized in every respect as that of a larger firm. But, it may be argued, the same will be true of a large firm whose size is not an integral multiple of the smallest optimal size firm. A little reflection shows, however, that a large firm will have an advantage in this regard. Suppose that x is the minimum optimal size. Now consider a firm whose size is $\frac{1}{2}x$ as against one whose size is, say, $3\frac{1}{2}x$. Administrative costs per unit are higher for the "$\frac{1}{2}x$" firm than for the "x" firm. But the "$3\frac{1}{2}x$" firm does better than the "$\frac{1}{2}x$" firm, since the "$3\frac{1}{2}x$" firm can have three "plants" of size x and one "plant" of size $\frac{1}{2}x$, whose *average* administrative costs must be lower per unit than the costs of the "$\frac{1}{2}x$" firm.[7] Thus, the average large firm that deviated from optimal size (i.e., an integral multiple of the minimum optimal size) would have lower administrative costs per unit than the average small firm if both small and large firms were equally distributed around the minimum optimal size or its integral multiples.[8] Of course, *some* small firms (of size x, say) will do as well on this score as do many large firms.

---

[6] Of course, the expansion need not be limited to the addition of exact duplicates of the initial unit. Firms may, and frequently do, "cash in" on the exploratory costs of organization creation by expanding in the direction of complementary products, or closely related products, as long as these employ administrative setups similar to the initial one. In many cases the similarity of the added product to the initial one may be such as to require only slight modification of the administrative structure. For example, a firm producing automobiles may easily add a line of trucks, or a firm producing one type of electrical appliance may easily add another one, and so on. Such additions also enable the firm to "cash in" on the administrative excess capacities that may exist at the time. For example, adding a line of complementary commodities that doubles sales revenue need not require a doubling of the sales force. The same salesmen may, almost as easily, show both lines simultaneously. The important point here is that the initial overhead costs of creating a successful and expandable administrative unit are spread over the additional substructures, and unit costs are thus reduced.

Also, the firm growth hypothesis developed in Chapter 16 implies that larger firms are, on the average, more successful than smaller ones if for no other reason than the fact that the large firms are the successful survivors of a previous generation, the weak members of which have left the industry or stopped their growth.

It has been argued that up to some point it may pay the larger firm to specialize and centralize certain activities that could not be afforded by the smaller firm. From this point of view also we might expect the average large firm to have lower administrative costs than the small one.

---

[7] A similar argument will be elaborated and formalized in the next chapter under the rubric of "the corkscrew cost curve."

[8] It is, of course, possible for the large firms, because of greater administrative and other experience and therefore greater knowledge of and insight into what the minimum optimal scale and its integral multiples might be, to be more concentrated near a multiple integral of the minimum optimal scale than the small firm. See the next chapter for an elaboration of these ideas.

The smaller firms may include among their numbers stagnant firms. The stagnant firm would find it difficult to introduce innovations, including administrative ones, whereas the large growing firm could more readily afford to do so in some of its units. In other words, the large growing firm with a somewhat higher profit rate can afford the risk of administrative experiments in some units, which, if successful, it could then apply elsewhere, whereas the small firm, fighting for survival, would not want or perhaps feel able to afford such risks. This notion is based on the idea that there will be a smaller number of those for whom survival is a serious question among large successful firms than among small firms.[9]

---

[9] See Chapter 19 on this point.

19

Cost Curves, Growth, and the Long-Run Theory of the Firm

INTRODUCTION

According to some general observations, as well as a number of careful studies, both in the United States and elsewhere, the following conditions appear to be fairly widespread.[1] These are for the most part conditions that are not explained by existing theory: (1) In most industries firms come in a large variety of sizes. (2) Small firms seem to earn on the average smaller profits than large firms. (3) Bankruptcy rates are larger for the small firms. (4) However, some small firms earn as large profit rates as the largest firms do.

---

[1] See W. L. Crum, *The Age Structure of the Corporate System,* University of California Press, Berkeley, 1953, pp. 93 ff.; also M. M. Mehta, *Structure of Indian Industries,* Popular Book Depot, Bombay, 1955, pp. 109 ff. See the bibliography in Mehta's book for other studies that reveal similar facts, and especially the bibliography on pp. 8–10. See also chap. 12 in P. J. D. Wiles, *Price, Cost and Output,* Blackwell, Oxford, 1956. It seems to me that this provocative book has not received the attention that it deserves. For Japanese data on the same phenomena see "Capital Accumulation and Efficiency of Japanese Enterprises After the War," *Monthly Circular—Survey of Economic Conditions in Japan,* May, 1959, pp. 10–16.

We will try to show in this chapter how some of the ideas emphasized in the previous chapters suggest cost curves of a type that can shed some light on these facts. While it would be unjust to argue that existing microeconomic theories necessarily contradict any of these conditions it is true to say that the usual textbook version of price theory[2] has, at best, nothing in particular to offer on them.

If we assume a state of perfect competition, in which all firms have equal access to all factors of production at the same price, we will find it especially hard to explain such situations. That is, under strict perfect competition all firms face the same long-run U-shaped cost curve, and therefore we would expect that in equilibrium all firms would be of the same size. Also, if firms started out at different sizes, we might expect that bankruptcy rates and losses would be greatest among the smallest firms and among the largest—that is, among those firms that deviated most from the optimal size—and we might normally expect the optimal size to be in between either extreme. However, the large firms do not have, on the average, as low profit rates or as high bankruptcy rates as the very small firms. In any event, we would not expect both small and large firms to survive simultaneously in the perfect competition case. But, of course, we need not assume that the strict theory of perfect competition is applicable here. We return, then, to our previous assertion that very little has been deduced from conventional theory in connection with the facts we have mentioned.[3]

In any case, if small firms are not presumed to be of optimal size, then it is certainly surprising that some small firms make as large profits as any firm in the industry; i.e., since the small firms would be deviants from optimality they would not be on the lowest point of their average cost curve and therefore we would normally expect them to have lower profits.

---

[2] Cf., e.g., one of the best of these texts, Alfred W. Stonier and Douglas C. Hague, *A Textbook of Economic Theory,* Longmans, Green and Company, London, 1954.

[3] Cf. the similar view in H. A. Simon and C. P. Bonini, "The Size Distribution of Business Firms," *American Economic Review,* September, 1958. See also the interesting discussion of many of these phenomena in chaps. 4 and 5 of William Baumol's *Business Behavior, Value, and Growth,* The Macmillan Company, New York, 1959.

These remarks are intended primarily to provide motivation for the following discussion. Our aim here is not to argue or to pinpoint the shortcomings of existing theory—even if the last could be characterized unambiguously. (No doubt many exceedingly useful models could be developed on the basis of the existing body of doctrine.) Rather, it is positive in spirit. We want to examine the potential explanatory value of some of the notions developed in previous chapters in the hope of adding to the existing stock of analytical tools rather than supplanting any of them.

THE CORKSCREW COST CURVE

Emphasized in the previous chapters is the fact that most agents of production are to some degree indivisible; they come in fixed sizes or, at the very least, some sizes are for production purposes superior to others. In working out our combination of factors we must realize that some of these factors will have fixed maximum capacities. In addition, as argued in the last chapter, it may be desirable for a firm to consider the duplication of known successful organizational structures or substructures when considering larger and larger scales of operation, rather than risk the uncertainty of trying to create an entirely new structure at each scale of operation.

The cost curve that we shall derive will be different from the conventional one[4] simply because it takes into account the implications for the cost-scale relationship of the fixed capacity of some of the factors or combinations of factors of production.

margi

The nature of our basic cost curve is illustrated in Figure 27. We assume that in our production function there is at least one factor or more of fixed capacity. The fixed factor can be increased or decreased through duplication but not through division. That is to say, we can have one, two, or three units of the same factor, but

---

[4] See Chapter 4 above.

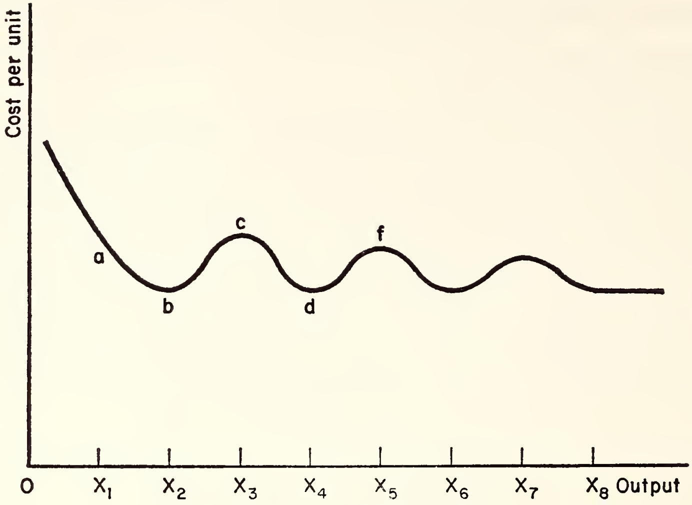


Figure 27.

not half a unit or any other fraction. In the first instance we would expect to experience increasing returns as we increase output, and at the same time increase the use to which we put the factor or factors with a fixed capacity. Beyond some point we may use the fixed factor so intensively (in the sense of crowding the use of the other factors with respect to the fixed factor) that we experience increasing costs with respect to increases in scale measured in terms of output. In Figure 27 we illustrate this phenomenon by the portion of the cost curve marked *abc*. (As usual the abscissa shows the units of output, and the ordinate, the cost per unit of output.)

Now according to the figure a low point in cost is reached when we produce Ox_2 units of output. For Ox_3 units average costs are higher because of the crowding phenomenon previously mentioned. Another way of looking at the matter is to say that with one fixed factor (or one set of fixed factors) there is an optimal combination of nonfixed factors which associated with the fixed factors will lead to lowest cost. For outputs on either side of this optimal combination average costs will naturally be higher.

Consider the production of twice as many units as Ox^2. The costs for producing that quantity need not be higher because we can simply duplicate the combination of factors, both fixed and variable, utilized to produce Ox_2 and obtain a combination of factors that can produce twice as many units, designated in Figure 27 by Ox_4. For the time being let us assume that such coördinating costs are always proportional to output. Later we will relax this assumption. But with it the average costs for producing Ox_4 units need be no higher than those for producing Ox_2. Assuming that there are no other economies of scale, unit costs will be exactly the same as those for producing Ox_2.

Now let us consider the cost of producing an amount in between Ox_2 and Ox_4, say Ox_3. Ox_3 is exactly three times Ox_1. By using three sets of the bundle of factors employed to produce Ox_1 we can obtain a set of factors capable of producing Ox_3. Under this procedure the cost of producing Ox_3 certainly need be no higher than the cost of producing Ox_1. But the important point to observe at this juncture is that under such circumstances there will always exist another combination of factors that will enable us to produce Ox_3 at a lower average cost than that required to produce Ox_1. For example, we may choose the bundle of factors used to produce Ox_2 and add to that the bundle of factors used to produce Ox_1. Both bundles will certainly be capable of producing an output of Ox_3, but their average cost per unit will certainly be lower than that required to produce Ox_1 alone, since we average in the lower unit cost of production obtained when we produce Ox_2. Thus, we have now obtained the curve *abcd* which has certain characteristics. The production of outputs that are exact multiples of Ox_2 will always be produced at lowest cost. The production of outputs that are not exact multiples of Ox_2 can be produced only at higher than minimum cost. However, *the less the degree to which noninteger multiples of* Ox_2 *require more than minimum cost, the greater the output.* We have already seen this to be true in considering the production of one-half the optimal amount (Ox_2) as against the production of one and one-half times the optimal amount.

We could repeat the procedure we have just gone through by beginning at the point *c* on the curve in Figure 27 and proceeding to the point *f*. We obtain similar results. An output of Ox_6 can now be produced at minimum average cost, while an output of Ox_5 can be produced at a cost lower than that required to produce Ox_3. Clearly, in producing Ox_5, when we add in the minimum cost required to produce Ox_4 and average it with that to produce Ox_1 (which will yield the output Ox_5), we get a lower cost per unit than that required when we produce Ox_3. Of course, for some noninteger multiples of Ox_2 there may be combinations of factors other than the integer multiples of factors used to produce Ox_2 and the remainder (by adding a unit of the size necessary that is smaller than Ox_2) that lead to still lower costs. In any event, the general shape of the curve is likely to be the same. Through this procedure we obtain a corkscrew cost curve—one in which all multiples of the optimal combination of factors are illustrated by troughs in the curve at a minimal cost level, but the successive peaks in the curve are each lower than the preceding one.[5]

The logic of the corkscrew cost curve can be easily seen by considering any two successive waves of the curve. The high point on a later wave must be lower than the high point on an earlier one. The reason is that no matter what the high point on the earlier wave is, it is always possible to produce the output associated with the crest of the later wave by producing the output associated with the previous high point plus that output (equal to Ox_2) associated with an optimal bundle of factors. The larger output must be associated with a smaller cost, since we average into the cost picture the minimal unit cost obtainable by producing the difference in the two outputs at minimal cost.

Two other influences must be taken into account. With multiple fixed factors it is possible, on the one hand, to run into decreasing returns with respect to coördination because some managerial capacities may be inaugmentable or relatively inelastic. On the other hand, as we increase scale it may pay to specialize one of the coordinative activities and achieve further decreasing costs thereby. Both of these influences would manifest themselves by an upward or downward tilt of the corkscrew cost curve. The two influences work in opposite directions. For the time being we shall assume that they cancel each other out. Later on, we shall consider the predominance of one or the other as a possibility and examine its implications for our analysis.

---

[5] An example that fits the corkscrew cost curve hypothesis very well is found in Sherman Maisel's interesting study of the San Francisco Bay Area housing industry, *House Building in Transition*, University of California Press, Berkeley, 1953. Cf. especially chap. VIII. On p. 196 Maisel reports that "the most effective amount of specialization is reached when there are between 50 and 100 men on the project. After this point, the same organization is simply repeated as many times as necessary to reach any desired level of output."

SOME EXPLANATIONS—FIRM SIZE VARIABILITY

We now apply the corkscrew cost curve to some of the facts mentioned at the beginning of the chapter. The corkscrew cost curve and the firm growth hypothesis (Chapter 16) can be used as a basis for the construction of models that imply the existence, and in some cases the persistence, of a variety of firm sizes in the same industry. The firm growth hypothesis establishes the incentive and logic for firms to start at a small size, while the corkscrew cost curve hypothesis establishes a set of conditions that enables the small firm, if properly managed, to compete successfully with the large ones. Some of these models will be briefly sketched in order to indicate the potential explanatory use to which our two hypotheses can be put. Of course, not all possible models of this type will be developed in the few pages that follow.

In many industries firms come in a large variety of sizes. Also, in many industries the variety of sizes persists for a long time.[6] In the short run we can explain the existence of size variety simply by arguing that the firms are in different states of disequilibrium. We may impose the proposition developed in our previous chapter that, given incomplete knowledge about the optimum size, it would not be the best strategy for a firm to try to begin at an optimum scale. Size variety is thus explained on the basis that since different firms began at different points, they are at different stages in their growth

---

[6] See the statistics presented in Wiles, *op. cit.*, pp. 219–222. Most of the figures come from various T.N.E.C. monographs.

[illegible]
ability to respond to changes.

paths. But under such circumstances we would not expect the distribution of firm sizes to persist. Indeed, we would expect more and more firms to reach the optimum size and stay there. In time, as more and more firms reached their optimum size, it would become easier for new firms to guess what the optimum size actually is, and as a consequence more and more firms would attempt to begin operations at that size. While the influences we have mentioned undoubtedly play a role, we need not limit ourselves to such an explanation.

Another possibility is to assume that not all factors of production are equally available to all firms. For instance, suppose that each firm monopolizes a different factor with a limited capacity. For each monopolized factor there may be a different optimal combination of other factors to produce the commodity, and therefore a different firm size may be optimal for each commodity. The owners of the monopolized factors would, in competitive equilibrium, each receive a rent that would equalize the unit costs of production for the different firms. Firm sizes could certainly vary under such circumstances. While this case too is a possibility, and while it is consistent with firm size variety, it is not a situation that seems to be important and general in most industries.[7] Therefore, other explanations are not otiose.

The main point of this section is to illustrate the desirability of invoking *both* the firm growth hypothesis and the corkscrew cost curve hypothesis simultaneously in order to obtain models that can rationalize the facts we are interested in. Consider some of the implications of the firm growth hypothesis without employing at the same time the corkscrew cost curve hypothesis. Suppose that the long-run cost curve that is in effect is the U-shaped cost curve usually employed by economists, and that it is the same for all firms in the industry. There are two main possibilities. We can assume at the outset that all firms begin simultaneously, or we can assume that they begin at different points in time. If they all began simultaneously and grew at the same rate, then we would expect them all to be of the same size and to end up at the same size. If, on the other hand, they grew at different rates we would expect the more rapidly growing firms to reach their optimum size first and to have lower costs than the others. Prices would then fall in the industry through the force of competition, and the slowest-growing firms would as a consequence be forced out of business. But as more and more of the larger firms reached optimum size they would cease their growth. In addition, as time passed, a larger and larger proportion of the industry would be made up of firms of optimum size, and eventually all the firms in the industry would be of optimum size—and, of course, of the same size. Thus we see that, at least in this case, an invocation of the growth path hypothesis alone does not, in the long run, explain the persistence of different size firms within the industry.

---

[7] This factor may be of significance in countries and in industries in which, for noneconomic reasons, firms are family owned and managed. In that case the differential managerial capacities found in different families will determine the sizes of various firms.

Now let us proceed as before, but with the alteration that we no longer assume that firms all begin at the same time. In other words, we take into account the phenomenon of new entrants into the industry. But, and this is quite important, we need not assume that entry remains equally easy under all circumstances. Once again consider first growth at a constant rate. The firms that start earlier will reach the optimum size sooner. As more of them do so the gap between the costs of the early entrants and the cost per unit of the later entrants widens. As this occurs, entry into the industry becomes less and less attractive. Simultaneously, more and more of the early entrants reach optimal size and cease their growth. What develops is that a growing proportion of firms become firms of similar size, while entry gradually becomes so difficult that new firms are a negligible proportion of those already at optimum size. Certainly the ratio of entrants to optimum size firms in terms of output capacity becomes smaller and smaller, and at the limit only the optimum size firms, all of equal size, remain in the industry.

We now have to consider the variation on the previous analysis under which the firms which enter the industry, and which do not fail, proceed along their growth path at different rates. The outcome here will be seen to be more or less the same. Even though the firms proceed at different rates to the optimal size, once they reach it they remain at that point. Prices in the industry fall as more and more firms reach optimum and the gap between optimal costs and the cost of the slower growers or the new entrants widens. Continuing the logic of the previous paragraphs, we see that eventually the remaining firms will, as a limit, all end up at an equal and optimum size. Clearly, the mere use of the growth hypothesis does not in the cases considered lead to an explanation of size variation within the industry *in the long run.*

It is worth noting that the imposition of the corkscrew cost curve without the use of the growth hypothesis can, under some restrictive circumstances, lead to long-run firm size variation. This would be true where all the troughs on the corkscrew cost curve are at the same level—i.e., where all minimum cost points are equal—and when the cost curves are the same for all firms. In that case there would be no special advantage for a firm to be small rather than large, *assuming that it has such perfect knowledge of cost conditions that it could always operate at the trough.*[8] But consider what happens under some slight variations of these rather strict assumptions. Suppose that the low points on the waves get lower and lower and that beyond some point they increase. In that case there will exist some low cost point which is lower than all others, and as a consequence firm sizes will eventually be concentrated around the output associated with the *minimum minimorum* cost point.

cool! now we have a U-shaped corkscrew.

Also, we might assume that the low points on successive waves are successively lower.[9] In that case all firms would begin at an indefinitely large size. Indeed, one cannot even say what that size would be. All we can say is that the larger the better. Alternately, if we assume that successive low points are higher than each other, the firms will all start at the low point of the first wave and stay there.

---

[8] We assume here that firms want to maximize profits. Should they want to maximize revenue subject to a minimum profit constraint, they would want to be indefinitely large under these circumstances.

[9] At least up to a point this is likely to be the case in many industries. Cf. Maisel, *op. cit.*, chap. 8. Indeed, Wiles argues that this is the normal case and the empirical evidence appears to justify this position. Cf. Wiles, *op. cit.*, chap. 12.

However, even where the low points of successive waves are at the same level it would be to the advantage of a firm to be larger rather than smaller if it took into account flexibility of output in the neighborhood of the low point—that is to say, if the firm took into account what would happen to its cost per unit if it had to produce a somewhat larger or a somewhat smaller than optimum output occasionally. The larger the firm the more shallow its cost wave and the less the increase in cost for any deviation from optimal output. If this were the only consideration, then firms would all want to be as large as possible.

Now let us combine our two hypotheses. At first we assume that all the low points on the successive waves are at the same level. Given our firm growth hypothesis, all firms will start at a relatively small size. They will have an incentive to grow because of the flexibility advantage of a larger size.[10] That is, the larger the size, the less the penalty in terms of profit reduction that can arise because of occasional nonoptimal output. However, with a corkscrew cost curve small firms do have a chance to survive if they pick the optimal output and do not deviate too much from it.

Now suppose that the firms start at the same size but at different times. For the time being assume that they all grow at the same rate. The early starters get larger and larger, but as they do so new entrants come into the industry. If there is a rapid enough expansion of demand for the product, some of the firms become indefinitely large, new firms enter at no disadvantage, and we have a picture of a steadily expanding industry with firms of different sizes persisting through time. It is worth noting that the rate of expansion implied by this model is an increasing rate of expansion of output, for not only do the existing firms expand at a constant rate, but each period new firms enter and in turn expand at a constant rate. But suppose that demand for the product does not expand at a growing rate. In that case a potential inconsistency enters the picture, and we cannot expect the industry to be able to maintain its rate of expansion of output.

---

[10] They may also want to grow for other reasons. See Chapter 16 above; also cf. Baumol, *op. cit.*, chap. VI.

In actual practice, of course, such inconsistencies are solved by bankruptcies and exits from the industry, and by a reduction in the rate of entry. There is no reason why we should not take account of these factors in our model. For the time being we will merely assume that for firms of a given age a certain proportion of them will leave the industry because of losses or less than anticipated profits. Up to a point, the younger the firm, the greater the chance of its leaving the industry.[11] In the next section we will argue this last point on the basis of other aspects of our model, but at present we assume it to be true for convenience in exposition. In addition, if there is a maximum age that any firm can reach, we can develop a model similar to that used in population problems. Despite the fact that our analogy may sound rather artificial at first, the reader is asked to bear with us for a few pages as we develop the argument.

We start with a given rate of demand expansion that is equal to the rate at which firms already in the industry expand their output. Now, given the demand situation, the inducement to enter will depend on the number and size of firms in the industry. Thus, the number of new firms will depend on the number of firms in each age-size group in the industry. This is clearly similar to birth rates, which depend on the ages of those already in the population. Also, at each age a certain proportion of the firms leave the industry, and this is equivalent to age-specific mortality rates.

Before proceeding further we shall look for a moment at the maximum age assumption. At first glance it may appear entirely unrealistic because there is no physical reason why a firm, especially a corporation, cannot persist in perpetuity.[12] Yet as a matter of observation, at least some old large firms do go into bankruptcy or leave an industry in some other manner. There is no denying the fact, of course, that some very old firms appear in every respect to be healthy, going concerns. However, the majority of these firms are not really at present in the industry in which they started. From this point of view they may be said to be no longer living with respect to the initial industry. Thus, it would appear that for some industries at least the maximum age assumption is not an unduly drastic abstraction from reality.

---

[11] This generalization appears to be well established empirically. Firms seem to be like children; the first five years are the most dangerous, and the younger they are, the higher the mortality rate. Cf. Wiles, *op. cit.*, pp. 219–223; also Crum, *op. cit.*

[12] However, in fact the mortality rate of corporations is not much lower than that of individual proprietorships. Cf. Wiles, *op. cit.*, p. 222.

Given the postulates stated in the last two paragraphs, we can now apply a theorem developed by Alfred J. Lotka[13] which he applied to problems of human population. Lotka's theorem states that, if entries into a population are a given proportion of those in each age group, no matter what age distribution we begin with, the population will eventually settle down into a stable age distribution. In other words, if age-specific birth rates (entry rates) and age-specific death rates (exit rates) are constant, then, as a limit, the population will approach a stable age distribution. The proportion of members in each age group will be constant although the population as a whole may be either constant, growing, or declining.

Let us see what happens if we apply these ideas to the problem at hand. We have assumed previously that all firms that do not leave the industry grow at a constant rate.[14] Given the age distributions of the firms in the industry, then, we know its implied size distribution. We also assume that of the firms within any age group a certain proportion will leave the industry during a period. Of course, this proportion will be greater for younger firms and up to a point less for older firms. If this proportion in each age group remains the same period after period, the analogy to constant age-specific mortality rates is exact.

---

[13] See Louis I. Dublin and Alfred J. Lotka, "On the True Rate of Natural Increase," *Journal of the American Statistical Association*, September, 1925. For a more extended treatment see Alfred J. Lotka, *Théorie Analytique des Associations Biologiques*, Hermann et Cie., Paris, 1939.

[14] While firms in fact do not all grow at a constant rate, it does appear to be true that on the average the larger firms are the older firms. See Wiles, *op. cit.*, for some of the statistical evidence. In the pages that follow we shall relax the assumption of strict constant rate of growth for all surviving firms.

Now, what determines the number of firms that enter the industry? Three factors: (1) the rate at which demand expands in relation to the expansion of the industry by firms already in it; (2) the number of firms in the industry; (3) the age and therefore the size of the various firms in the industry. The assumption that all firms grow at a constant rate implies that the size of the industry depends on the number of firms and their age distribution. For the time being we consider the case under which demand expands at a constant rate. Under these circumstances profits will depend on the rate at which demand expands as against the rate at which the industry expands. The greater the gap between the two, the higher the profit level and the greater the inducement for new firms to enter. But, in addition, we must consider the relative efficiency of the different firms. If, on the average, a higher proportion of larger firms achieve a certain degree of efficiency than smaller firms (for reasons that will appear in a later section), then the degree of efficiency in the industry will depend on the age distribution. In other words, new firms will be induced to enter to a greater extent if the industry contains mostly young firms than if it contains old, experienced, and on the average more efficient firms. But this statement implies that the number of births of new firms is a function of the age distribution. This is analogous to the condition necessary for the application of the Lotka theorem—namely, that of *age-specific* birth rates which are constant through time.

The result is that, given our assumptions, the industry in question would eventually be populated by firms, some entering and some leaving, which have a constant age distribution and a constant size distribution. Thus, we have a model consistent with the persistence of firms of different sizes within the industry.

The reader might argue, perhaps, that there are artificial elements in the above model. This is, of course, likely to be true of all models and it is certainly true of this one. Therefore, consider the possibility of relaxing some of our assumptions. For example, the size of a firm is not an exact function of its age, just as the size of an individual is not an exact function of his age. We may assume, instead, that for every age there is a given distribution of sizes,[15] but that this distribution remains constant given the age group. It will, therefore, follow that if the proportions of the population in each age are kept constant the proportions of the population in each size will also remain constant. Thus the relaxation of the strict size-age relationship *need not* change (but of course it may change) the general picture.

Now consider relaxing the assumption of a maximum age limit. If the inducement to enter declines as the proportion of older firms in the industry increases, and if some firms live on perpetually and grow perpetually, then, of course, the age and size distribution within the industry will not remain constant. Instead there will be a bias in the trend toward older firms. But as long as there is an inducement for *some* new entrants into the industry a variety of firm sizes and firm ages will continue.

On the other hand, it might be possible to construct a model in which the reverse tendency is illustrated. For example, suppose that as firms get very large their rate of growth declines—that somehow, beyond some point, their management structure becomes so unwieldy they find it difficult to make the necessary changes required for the maintenance of the growth rate. In that case as the older firms increase in age they may grow less rapidly than the expansion in demand requires, and thus create an increased incentive for new firms to enter the industry. Under such conditions the bias could be toward more and more younger firms if the expansion of total demand is rapid enough. Clearly, the outcome in such models will depend on the relative rates of growth of demand, the growth of firms at various ages, and the inducement to enter as a function of demand and firm growth.

It should be clear from the previous paragraphs that we have been manipulating a number of variables and relations. Quite a few models could be built on the basis of such variables and relations, some of them different from the ones sketched above, that would lead to various patterns of size-age distributions over time.

---

[15] But we retain the assumption that on the average older firms are larger than younger ones.

Some of these would imply a stable size pattern as a limit, but others would imply perpetually changing age-size distributions. It is not in our interest, for present purposes, to elaborate all the possible models.

Enough has been said, however, to illustrate that with the assumption we have introduced we can easily construct models to explain, either as an equilibrium or as a disequilibrium phenomenon, the persistence of firms of a large variety of sizes within the industry. As already stated, some of these models would imply an eventual stable age-size distribution, others a growing bias toward larger firms, and still others a bias toward smaller firms. At this stage it is worth while to stress again the two important elements that enable us to build such models. These are, first, the growth hypothesis, which establishes the incentive and logic for firms to start at a small size, and second, the corkscrew cost curve hypothesis, which establishes a condition enabling the small firm, if properly managed, to compete successfully with the large one.

OPTIMAL VERSUS NONOPTIMAL MANAGEMENT

In the previous section we assumed that with a given scale of operations, costs are strictly determined. In actual practice this is usually not the case. Any *given* output may be produced at higher or lower costs depending on how well production is managed. In other words, we need not assume that given the output every firm can produce that output at minimal cost. Only a small proportion of the best-managed firms can do so. For any given output we might expect a normal distribution of degrees of efficiency of production. Figure 28 illustrates the situation by showing four corkscrew cost curves. The bottom one (curve *A*) shows the absolute minimum cost, achieved only by the most efficient firms in the industry. Curves *B* and *C* show the bounds within which the majority, say 90 percent, of the firms will be producing. Most firms will have costs somewhere between the cost levels of *B* and *C*, such as curve *M* in the figure. In other words, the median firm may expect to

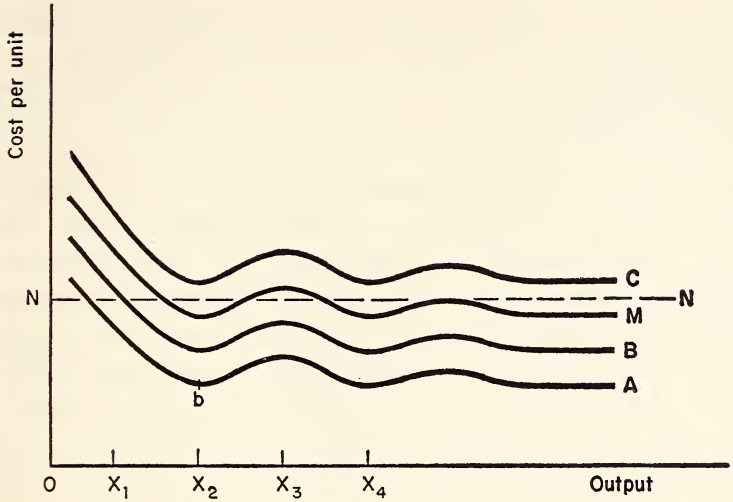


Figure 28.

produce on the cost curve *M*, while an especially efficient firm may get below *M* and a less efficient one will have costs above *M*.

Let us consider the first wave on our set of corkscrew cost curves. Only one point on that wave represents true minimum costs (the point *b* in the figure). There are clearly two reasons why a firm producing an output in the *neighborhood* of Ox_2 will fail to produce at minimum cost: (1) because the firm has not chosen the optimum output on that wave, or (2) because the firm is not managed with maximal efficiency, so that it is not producing the output, whether optimal or not, on the lowest cost curve.

With this addition to our model in mind we can now consider two other possibilities not considered in the last section: that of periodic increasing costs with respect to scale, and that of periodic decreasing costs with respect to scale. By periodic increasing costs we mean that the low points on successive waves are each higher than the last, and by periodic decreasing costs we mean that the low points on successive waves are each lower than the last.

The increasing cost case is not especially interesting. Under such circumstances we would expect firms to stop their scale of operations at the low point of the first wave and stay there. Then all firms would be the same size and would be relatively small.

The important case to look into is that of overall decreasing costs, either indefinitely decreasing or decreasing up to a point—in other words, the case under which for quite a few waves on our corkscrew cost curve we have periodic decreasing costs. If all firms were able to produce on the lowest cost curve (curve *A*), then clearly the advantage would be with the larger firm, and late starters would be forced to accept lower profits, or losses, until they grew as large as the largest firms. But the largest firms may be able to grow as rapidly as the smaller ones and as a consequence the late starters would always be at a competitive disadvantage. However, if for each output the majority of firms do not produce at minimal costs, it is quite possible for a small firm to do as well as a large one if it produces on a curve closer to the minimum cost curve than the large firm. Under such circumstances small firms could still enter the industry, and some of them would be able to compete successfully with the large ones and survive. Of course, only a small proportion (certainly less than half) of small firms would be able to do as well as the median large firm.

FIRM SIZE, PROFITS, AND BANKRUPTCY RATES

Continuing our analysis on the basis of the ideas developed in the last section, we now examine the relations between firm size, profits, losses, and bankruptcy rates. First, we consider the periodic constant cost case and then the periodic decreasing cost case.

Figure 28 again illustrates the situation in the periodic constant cost case. As before, curve *A* is the minimal periodic cost curve and curve *M* represents the average cost of the median firm for each output. (That is, at each output half of the firms will produce above *M* and half below it if there are no exits at that output.)

Now for each firm size (measured in terms of their output) various distributions of firms around the median firm are possible. For purposes of simplification we will assume that the distribution of firms of a given size in terms of their unit costs of production is symmetrical around the median.[16] Also, for each firm in a given size class we associate a unit cost of production. The distribution of costs around points on the curve *M* may be the same for all outputs or different for different outputs. If the distribution of costs is the same at all outputs, the curve *M* will be the same distance from curve *A* throughout. Again, larger firms, because of greater experience, may be, on the average, more efficient than smaller ones, in which case *M* and *A* approach each other with increases in output.[17] These are, of course, not the only possibilities, but they seem the reasonable ones to consider.

At first, for ease of exposition, consider the case in which *M* and *A* are equidistant throughout. Now the line *N* in Figure 28 indicates the equilibrium price level. That is, the price *ON* equates supply and demand. If demand is stable, *ON* is the price that induces a rate of entry of new firms, and a rate of growth of the existing firms, so that the increase in output achieved thereby is equal to the output loss arising from those firms that leave the industry at that price. If capital is in the same proportion to output for all firms, the relative profit rate will be in proportion to the difference between price and unit costs (profit per unit) for the various firms. It is possible that larger firms could use their capital stock more effectively than smaller firms and therefore their capital-output ratio would be lower. Thus larger firms could have the same profit rate at a lower profit per unit of output. But for present purposes we wish to ignore this variable and assume the same capital-output ratio for all firms.

Now the bankruptcy rate will depend on the loss rate. Of course, the connection between bankruptcies and losses is quite a complex one. The mere fact of loss does not determine bankruptcy. But it is not our problem here to evolve a detailed theory of bankruptcy determination. Generally speaking it seems reasonable to suppose that the greater the loss rate, the greater the bankruptcy rate, other things equal. For our purposes this generalization of what is the probable situation will suffice.

---

[16] Mehta's data for seven Indian industries suggests that this is really a fairly good assumption for each size class. See Mehta, *op. cit.*, the tables on pp. 130–146.

[17] Our two hypotheses would suggest that the dispersal around the mean should be greater for small firms than for large. Small firms have on the average had less experience and hence have acquired less knowledge of how to cut costs down toward the minimum for a given output. In addition, if we consider fairly wide groupings of firms, such as small, medium, and large, the smaller ones, for the same distribution of outputs around the optimal output on each wave, will have a greater dispersion of costs than the larger firms, since the cost waves of the latter are shallower. The statistical evidence appears to support this conclusion. See Wiles, *op. cit.*, pp. 233–246, and Crum, *op. cit.*, pp. 27 ff.

As we look at Figure 28 it becomes evident that large firms will suffer fewer bankruptcies than small ones. For example, the median firm producing an output of Ox_3 will incur losses since its cost level is above the price line. However, there will always be a larger firm producing at the crest of its cost wave which will have that crest below the price line since, as we have shown in previous sections, the crest of the successive waves must get smaller and smaller. Clearly, the magnitude of the losses will be larger for small firms operating at the crest of their cost waves than for large firms doing the same thing. If bankruptcies are positive monotonic functions of loss magnitudes, small firms operating at their crest will have a higher bankruptcy rate than large ones operating at their crest. Also, if small firms are no better than large ones in determining what their size ought to be, the bankruptcy rate will be higher for the small firms than for the large. In other words, if we assume that in each broad size class the distribution of firms around the optimal size on each cost wave is the same, the bankruptcy rate will be higher for the smaller firms.

Since the later waves are shallower than the earlier ones, profits must be higher on the average for the larger firms than for the smaller ones. This conclusion is entirely apart from the question of bankruptcies. But it follows immediately from the conclusion of the higher bankruptcy rate because even if small firms that do not go bankrupt earn the same profit rate on the average as the large ones, once the higher proportion of bankrupts is added in, the average profits rate must be lower.

We argued the above on the assumption that the curve M was equidistant from the curve A throughout—i.e., that the degree of efficiency achieved by the median large firm was no greater than that for the median small firm. Should we make the assumption that median large firms are more efficient from a managerial point of view, as we argued in the last chapter and as is suggested by some of the empirical evidence, then all of our conclusions are reinforced. The large firm will earn higher profits, not only because of the shape of the corkscrew cost curve, but because of greater relative managerial efficiency for its output.

It is worth considering for a moment why we might expect the median large firm to have greater managerial efficiency than the median small one. There are two closely connected considerations. From the firm growth hypothesis it follows that large firms have had more successful managerial experience. To the extent that managerial experience contributes to managerial proficiency, we would expect the larger firms to be more proficient than the smaller ones.[18] Along the same lines we must remember that the larger firms are, managerially speaking, a more select group than the smaller. The smaller ones contain a certain proportion of managerial incompetents who will at some stage be forced to leave the industry before they grow very much larger. The larger ones are by the same token the survivors of successful firms and are therefore more efficient on the average than those in the ancestor group.

Figure 28 shows, however, that some *small* firms do as well as any large firm. A firm producing Ox_2 on curve A will produce as efficiently as any firm whatsoever. There is no implicit reason why such a small firm cannot exist. Of course, highly efficient small firms will be a smaller proportion of their size group than would be the case for large firms.

Suppose we consider briefly the same picture but assume that the corkscrew cost curve is a periodically declining one. In that case, all of our conclusions except one will hold. The analysis would be exactly the same in all cases. We have already considered the fact that this situation may still permit small firms to enter, since efficiently managed small firms may still do better than inefficiently managed large ones. It follows, obviously, that small firms will earn, on the average, smaller profits than large ones. Similarly, the bankruptcy argument could be repeated as before. The argument would be strengthened by the fact that falling periodic costs would give an equally well-managed large firm a cost advantage over the equally well-managed small one, producing on the same relative point on its cost wave. However, our conclusions would no longer be exactly parallel in one respect: the optimally managed small firm can no longer do as well as the optimally managed large firm whose size is on the trough of its cost wave.

---

[18] That efficiency is often related to past experience is brought out by the phenomenon of production "learning curves" emphasized by industrial engineers. See "The Learning Curve as a Production Tool," by Frank J. Andress, *Harvard Business Review,* January–February, 1954, pp. 87–97.

20

Notes on Organization and Oligopoly Behavior

NONPARAMETRIC DECISIONS AND OLIGOPOLY

Oligopolistic markets possess elements that are different from those found under competition, monopolistic competition, or pure monopoly. Briefly we might say that there are potentially "nonparametric"[1] elements in the situation, as the economist views it, that do not exist in other types of market structure.

We may distinguish between "parametric" and "nonparametric" decision-making. The essential nature of parametric decision-making is that one can, in principle, calculate the value of the consequences of selecting any one of the alternatives available. That is to say, the value of an alternative depends only on the nature of the choice and not on anybody else's action. In the nonparametric case the values of at least some of the variables will depend on the unpredictable reactions of one's competitors to the behavior of the firm in question. For example, the quantity that a duopolist can sell will depend not only on the price he sets but also on the price set by the competitor for the same commodity or for a close substitute. Price here is a nonparametric element. It is not a variable that is completely controlled by the firm. How should nonparametric elements be handled? This is the essential problem for the oligopolist. We shall discuss some elements of it in this chapter in the light of some of the points made earlier.

---

[1] A parameter in mathematics is a variable that for purposes of the problem at hand may be viewed as a constant. That is to say, it is an element which can take on different values but which is assumed to have a specific value for purposes of a particular problem.

Usually in economic discussion different market structures are defined in terms of the number of competitors within a market. Although this is not the only characteristic that is usually considered, it is a basic one. Thus, in perfect competition there are characteristically a great many small producers in the same industry. Oligopoly is usually viewed as a competition among relatively few producers. Yet a little reflection should indicate that the number of producers is not really the crucial element. As Wiles[2] has pointed out, the basic element of oligopoly is that under oligopoly a producer, in making his decision, has to consider the *personalities* of other producers. Under competition a producer does have to take into account what other producers will do, but he has to worry about them only as a group and not as individuals. Under monopoly there are, of course, no other producers about whom to be concerned. But under oligopoly what other producers do is significant, and what any individual large producer chooses to do is especially significant. The essential element in the oligopoly situation is whether there exist in the industry producers sufficiently large so that the actions of any one of them influence the payoff to any other single producer.[3] Thus, an industry in which there are, let us say, two large producers and a great many small ones is just as much of an oligopoly situation as one in which there are only three large

[2] P. J. D. Wiles, *Price, Cost and Output*, Blackwell, Oxford, 1956, chap. 6.

[3] From this point of view it is of interest to note that a lack of freedom of entry is not an essential aspect of oligopolistic markets even in the long run. The significant element is the existence of significantly large firms, and we saw in the last chapter that large firms may persist simultaneously with freedom of entry.

producers. What is of the essence here is whether there are producers sufficiently large so that their actions can be ignored by other producers only at their peril. (See Chapter 6.)

By contrast, under competition the personality of the entrepreneur is of no importance. Since the firm is small, no matter how peculiarly it chooses to behave, its activities can have little influence on the total market price or output and hence can have little effect on the destiny of any other firm in the industry.[4]

Looking at oligopoly as a situation in which the personalities of competitors are important sheds light on a significant element of the oligopoly problem. The difficulty of analyzing oligopolistic behavior is that all sorts of things can happen in oligopoly that cannot happen under other market forms. For instance, price wars are possible, competitors may engage in practices that attempt to drive other competitors out of the industry, some competitors may form coalitions at the expense of other competitors, etc. Under certain circumstances firms may find themselves engaged in a war of survival. The choice of strategies by firms under personal competition may lead to very wide price fluctuations. But although a great variety of erratic things *can* happen under oligopoly, in the usual situation very few do. Instead of price wars of survival we usually have truces under administered prices with considerable price stability.

If any of the things we have just listed happens, it is simply because one of the large firms in the industry wants it to happen. A firm's actions may be rational from the point of view of its economic objectives, or they may represent the caprice of the owner, or they may be accountable for only in terms of the specific personality characteristics of the owner or managers of the firm. It is in this sense that personality is significant. For example, a firm may start a price war in order to drive another firm out of the industry and thereby increase its long-run profits. Or it may start a price war because the owner seeks the power and prestige he believes to be attached to being the dominant firm of the industry, although such a venture may so endanger his own position in the long run that it would not be considered wise from a purely economic viewpoint. Once again, personality is important, and the activities of large firms may, to some extent, be viewed as extensions of the personalities of the owners or managers.

---

[4] Also, under competition, capricious behavior on the part of the entrepreneur will normally lead to nonoptional behavior, which in turn may lead to losses and eventually bankruptcy. If the industry approaches long-run competitive equilibrium, the firms that survive in the long run are those that do not behave capriciously. See A. A. Alchian, "Uncertainty, Evolution, and Economic Theory," *Journal of Political Economy,* June, 1950.

We have suggested that under oligopoly a great many things can happen that do not happen under competition. However, we do not intend to give the impression here that simply because a great many things can happen they in fact do happen normally. The danger to a firm in being in an oligopolistic industry is that it faces a *potentially* nonparametric decision-making problem. But this does not mean that, under what the managers of the firm would look upon as normal conditions, actual day-to-day decision-making is really nonparametric. To be more specific, a firm in an oligopolistic industry does face the risk of finding itself in a price war, in a war of survival, or in a situation in which his competitors form a coalition against him.[5] These are the potential dangers, but the word "potential" is to be stressed. There is usually an organizational alternative that will minimize or even eliminate these dangers. In other words, the firms in the industry can usually turn a potentially nonparametric situation into a parametric one. Part of the problem of understanding oligopoly behavior is understanding how this is done, and how a parametric state of affairs is enforced, either tacitly or overtly. Hence, we shall consider here (1) some ways in which oligopolists can handle the *potentially* nonparametric element in the oligopoly situation, (2) the relationship between long-run costs and the development of oligopolistic market structure, and (3) some of the organizational elements in the development of oligopolies.

[5] These are the sort of problems that are considered in the theory of games. See J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, 1953. See also M. Shubik, *Strategy and Market Structure*, John Wiley and Sons, New York, 1959.

ORGANIZATIONAL ASPECTS OF THE ENTRY PROBLEM

We shall approach our problem indirectly. We first consider some of the organizational requirements for a perfectly competitive market and then show how deviations from these conditions may lead to the creation of an oligopolistic market. We shall concentrate our attention on the question of entry under conditions of long-run constant costs with respect to scale for firms *already established* in the industry.

For an industry to remain perfectly competitive in the economist's sense of the term it is necessary that each firm, and each potential firm, have equal and free access to four areas in which the firm must operate: (1) the factor markets, (2) the commodity markets, (3) the production area, and (4) the finance market. Of course, such a subdivision of the firm's activities is arbitrary but useful for present purposes.

The consideration of equality of access of new firms to various markets is important because a lack of equal access implies an advantage to firms already in the industry. If the advantage is considerable, then oligopoly is likely to develop, since in the normal course of events some of the firms in the industry go bankrupt or quit the industry for other reasons. Thus room is left for the expansion of those firms that are already in the industry, which become progressively larger. In the case of equal access some of this room is occupied by new firms entering the industry, which in turn reduce the industry-wide significance of the surviving firms.[6] Of course, the extent to which access to various markets is unequal, and the extent to which it remains so after a firm enters the industry, will determine whether such inequalities become a significant deterrent to entry. Clearly, there is some degree of inequality which will operate as a deterrent.

---

[6] Whether or not oligopoly develops under these circumstances depends on the relative rates at which established firms expand as against the rate at which new firms enter. In any event, ease of entry at the same unit costs as established firms have is important because it makes it impossible for the established firms to charge noncompetitive prices in the long run.

Now, what is meant by free and equal access? Consider this matter with respect to access to the market for factors of production. First, for a firm to operate it must have *information* as to factor availabilities. That is to say, it must somehow be able to find out where the various factors of production, such as labor and raw materials, can be purchased. Second, it must be able to communicate with the sellers of factors; hence the *channels of communication* must be open to it to the same degree as they are to any other firm. Third, it must be able to make contracts with such sellers on the same terms as any other firm. Fourth, the sellers of factors must be willing to perform in accordance with their contract to the same extent as they are for any other firm. Fifth, should there be a lack of performance, the firm must have legal recourse to the same extent as any other firm. And sixth, all of this must be on a continuing basis to the same extent as it is for any other firm. That is, the firm must feel not only that it could do all these things once or twice but that it can do them indefinitely as long as other firms are able to make similar arrangements.

Another way of looking at the matter is to say that the new firm has to establish itself in a variety of roles vis-à-vis other firms. In this respect it has to create certain viable role relationships between it and (1) potential buyers of the commodity it sells, (2) sellers of raw materials or other nonlabor inputs, (3) suppliers of finance, and (4) suppliers of labor. One of the main relations to be established is that of mutual confidence between the new firm and other firms or market entities with whom it has to deal. In other words, the other firms have to feel the same degree of confidence in the new firm with respect to all potential interdependent activities between them as they do with any of the competitors of the new firm.

Thus a perfectly competitive market cannot exist simply because there are many small firms wishing to operate in that market; it requires organizational means to enforce continued equality of treatment for all existing firms, and especially for any potential firm that chooses to enter this particular industry. In other words, the firm operates within an organizational matrix larger than itself in which individual firms play certain recognized roles. Some firms or organizational entities must provide information and means of communication. Others must facilitate the making and maintenance of contractual arrangements. Still others must stand ready to perform in various ways. Finally, there must be a legal system and a set of organizational entities through which it operates so that firms know they have recourse in cases of nonperformance. Looking at the matter in this way we see that in back of the concept of a competitive industry we must have in mind the simultaneous existence of numerous other industries organized in such a way as to give equal access and equal treatment to any firm in the industry in question.[7]

Now an industry may conceivably develop *some* of the characteristics of a competitive market through mores and practices that arise within the industry rather than through the existence and utilization of organizational entities outside of the industry.[8] Firms may believe that it is to their advantage to develop a reputation of equality of treatment to all comers. This is especially likely to be the case with respect to commodities that have to be purchased at frequent intervals. The same is often true with respect to quality and other aspects of service to customers. Similarly, price and quality information may be posted conspicuously if it is believed that customers desire it. We cannot go into detail, of course, as to how such mores do develop. But once developed they may persist because the role relations established between firms are based on the mutual expectation of their persistence. When such mores spread to a great many industries, we have some of the conditions for equal treatment to all buyers. This would appear to suggest that free and equal entry can occur without any organizational means to enforce it. But in many cases (perhaps most) it is unlikely to occur with respect to *all* of the activities necessary to assure free and equal entry.

---

[7] Inequalities of access or treatment will not of necessity lead to oligopoly; in some cases they may lead to monopoly and in others to monopolistic competition.

[8] For example, the stock markets and the primary commodity exchanges, in developed countries.

There are some elements that depend on learning through experience, and usually the development of business mores and practices does not counteract the experience requirements. Consider the matter of finance. Not everyone who is able to run a firm can obtain the necessary commercial backing to do so. In theory, such backing should be available to every competent entrepreneur. But clearly firms who are already in the industry and who are doing reasonably well have an advantage over an outsider. The outsider faces a dilemma. In order to get financial backing to enter the industry he must demonstrate his competence, but he can only do so by being in the industry, or in some allied industry, in which case he would not need the financial backing in order to enter it. A similar argument could be made with respect to certain types of skilled personnel whom one might want to attract to an emerging firm. A trained person would usually wish to attach himself to a proved firm as against an unproved one. The same is true of production know-how. All of the details as to how to produce a commodity within the content of the firm cannot be determined in advance on the basis of some formula. Experience usually helps. It takes time to get the bugs out of new production procedures.

Now, if there were a learning curve that was definite and known to everyone equally well, experience (learning) would be an investment cost of entering the industry and everyone would have to bear a similar experience cost whenever he entered. But usually there is a large uncertainty element connected with experience, learning, etc., that gives the established firm an advantage. First, let us recall that the established firms are the *successful* survivors. Many of those who started earlier failed to survive. (See Chapter 19.) In other words, these firms have established their capacity to perform to investors, employees, suppliers, and so on. For example, established firms have proved their capacity to create and maintain a workable organizational structure. Surely this is not the case for the unproved entrant. Hence, for investors, suppliers, etc., they are not equal risks. Second, the established firm has already created for itself a workable role in the various markets and areas in which it has to operate, especially with respect to obtaining information, maintaining contacts, living up to contractual arrangements, and so on. The new firm has not done so, nor is it known that it could do so successfully. If there is room for more capacity in the industry, surely the proved performer will find it easier to get the necessary finance, credit, supplies, etc., to fill the gap rather than the new firm.[9]

On the other hand, there is the possibility that the new firm can imitate the organizational procedures and production techniques of the established firms at less cost than it took the established firms to develop these procedures and techniques *de novo*. In this case the outcome will depend on whether the imitation advantage can offset the disadvantages of not having proved one's abilities at the outset.

We are not suggesting that new firms cannot enter an industry, but in many cases they cannot do so on equal terms with existing firms.

Indeed, I suspect that entry takes place not because the potential entrepreneur can enter on equal terms but primarily because of some of the inequalities in the situation. For example, a potential entrepreneur may be an innovator with respect to some element in the production process. Hence he may be able to enter the industry precisely because he feels he has some advantage over existing firms —an advantage that counteracts some of the disadvantages he faces. Similarly, people with superior entrepreneurial talent may flow to industries and organize new firms in areas in which such high-quality talents have not operated previously. Or some people may enter an industry on the basis of noneconomic motives. They may want to be their own boss, or like to deal with the specific commodity involved (e.g., books and the opening of book shops), and so on.

Firms will be deterred or find it impossible to enter an industry if they cannot gain access to one of the markets or areas necessary for their activities, or if existing firms have distinct advantages over newcomers. Looking at the matter from the other end we can see that existing firms in an industry can deter entry if they in turn control one of the necessary factors for a new firm's activities, or if matters are so arranged that they have a distinct advantage in competition with new firms. We consider here some of the organizational means through which such situations may arise.

---

[9] Established firms in other lines entering a new industry do not face these problems to the same extent as entirely new firms; this may explain why entry is so often via old firms going into new product areas.

A firm may restrict entry by the extension of ownership in crucial areas. For example, a firm may purchase some of the supplying firms to the industry in question. If it purchases enough of them so that it controls the supply of one or more factors, then clearly it could restrict access to such factors by other firms or it could discriminate against other firms. Now, it may not wish to discriminate against other firms that already exist in the produce market, because the existing firms could compete with it in other ways and they therefore have some means to retaliate against discriminatory practices. However, the same power does not exist for a potentially new firm if the latter cannot get started in the industry for lack of access to a factor of production. In addition, the existing firms in the industry are likely to see the danger of potential discrimination against any one of them and, therefore, they will be willing to cooperate with each other, if necessary, against the discriminating firm. But they would all agree to deter the entry of new firms and hence they have no objection to such practices against outsiders. Indeed, if between them they controlled the supply of some factor, they could readily agree on nondiscrimination in the sale of that factor to each other but have it in their common interest to agree to discriminate against newcomers in order to deter entry.

The extension of ownership as a means to get control of production factors and to restrict access to potential entrants can be employed in a number of directions. For example, attempts may be made to extend ownership to sales channels, to market information, to production knowledge, patent rights, trademarks, and so on, and especially to sources of finance. We need not go into great detail here but it should be clear that with ownership usually goes control, and with control the means to restrict access to potential entrants.

Other means of achieving control are through contractual arrangements, coalitions, gentlemen's agreements, informal understandings, interlocking directorates, and mergers. All of these means differ essentially for our purposes in the degree of formality of the agreement and in the extent of recourse in the event of noncompliance.

For example, an agreement for the sole use of a patent will have legal standing even though it gives the parties to the agreement a competitive advantage over any newcomer who could not get to use this particular patent. Sometimes an agreement may not involve legal recourse but it may enable one party to the agreement to have other forms of recourse. For instance, if a firm agrees to be the sole supplier of some factor of production to another firm and if the buyer is a relatively large firm, the seller may fear to break the contract, not being sure he can sell his output to other firms without the large firm's patronage. Furthermore, a reputation of noncompliance with contracts is not desirable for a supplying firm to have. Interlocking directorates may result in exceedingly informal agreements to discriminate against newcomers, but in this case the means of communication are so convenient that it would be easy to maintain such policies. Mergers involve the complete merging of control of separate entities so that it becomes a simple matter for the new entity to agree on policies that will discriminate against newcomers.

Thus far we have developed our analysis on the basis of constant unit costs with respect to scale. But we should consider the likelihood of increasing costs with respect to increases in the scale of operations and control. Let us look at the matter of extending ownership. Will the mere change in ownership change the costs per unit of the component firms that come under the new ownership? What is implied by the change of ownership for our purposes is that with this change firms that were previously competitors could now follow a unified price and output policy. It is important to observe that the same result could be achieved by an agreement between the component firms without any formal change in ownership. Would such an agreement lead to increasing costs? In most types of manufacturing and distribution it is difficult to see why the mere extension of a unified ownership to cover former competitors, or the use of other means to achieve unified price and output policies on the part of former competitors, should lead to increasing costs.

However, if we do assume increasing costs with an extension in the scale of operations (and by an extension in the scale of operations we mean any scheme that leads previously separate firms to come under a unified price and output policy), then, of course, long-run increasing costs would be a deterrent to the development of oligopoly. If the increasing costs were sufficiently precipitous and began at a relatively small output, the deterrent effect could be sufficient to retain the competitive structure in the industry. In addition, if there were no way to prevent entry at equal or close to equal terms, clearly the competitive structure would be reinforced by the fact that new firms could enter at low cost—that is, at a cost lower than that of relatively large firms.

DECREASING COSTS AND THE DEVELOPMENT OF OLIGOPOLY

In previous chapters we have developed models of industries in which large firms and small firms coexist. We now consider some slight variations of these models in which industries that are more clearly of an oligopolistic nature are developed—that is to say, industries in which the surviving firms are all large.

Suppose we start with a set of relatively small firms of roughly equal size. Let us make the following assumptions and see what happens to the structure of the industry over time. First, we assume that in the long run there are decreasing costs with respect to scale, *provided* a sufficiently long period of adjustment takes place in the process of expansion. Second, we assume that unit cost of output is a U-shaped function of the rate of expansion. This means that there is an optimum rate of expansion—one that leads to lowest costs at the scale of operation into which the firm has expanded. We consider briefly why this might be the case. Firms that expand too rapidly get into organizational areas much beyond their range of experience. As we have argued in previous chapters, organization is not a standardized commodity that can be obtained in all sizes ready for use. The less rapid the rate of expansion, the more the firm can use current knowledge and experience in the creation of a workable organizational structure. Otherwise the existing roles and the existing structure must be unusually flexible to be able to absorb the added organizational entities necessary in order to operate at the much larger scale. It takes time to discover and absorb the additional personnel and other resources that are required at a much larger scale. Also, since organization is not standardized and known for all sizes there is uncertainty connected with the construction of an organization very far outside the current size class. In other words, it is uncertain what the level of costs will be after the period of adjustment, and probably the greater the increase in size, the greater the uncertainty. A very large increase in size will usually involve the creation of excess capacities that the firm hopes it will be able to employ when it captures a larger size of the market via lower costs. But the greater the excess capacity, the greater the risk of not getting costs down far enough, as well as the greater the loss during the period of adjustment. As a result of these risks it may be easier to raise capital through normal channels at nonpremium rates for small jumps in size rather than very large ones. Also, for small expansions the firm can often use its own capital whereas for large ones it has to go to the capital market.

On the other hand, the rate of expansion can be too slow. There are often discontinuities in plant sizes and hence it may be costly to expand by little bits. Because of such discontinuities it may be expedient to expand in jumps but not in exceptionally large jumps. If the jumps are too small, the indivisibilities and discontinuities appear, as well as the inability to duplicate completely existing organizational substructures.[10] But if the jumps are too large, the problem arises of having to absorb large additions to the organization on the basis of inadequate experience and familiarity.

Let us now examine the effects of these assumptions on the development of the industry. Clearly, the optimally expanding firm will have a cost advantage. Firms that expand too slowly or too rapidly will be less efficient and face the risk of losses and failures. They have the disadvantage of not being able to reduce their costs as rapidly as can their competitors who do not expand quite so slowly. If the price in the industry is established on the basis of the costs of the most efficient firms, then the others will be forced out of the industry. In actual fact the price is likely to be established at a point somewhat higher than this, but nevertheless at least some of the high-cost firms will be forced out. As this procedure continues the optimally expanding firms will get a larger and larger share of the market. This is clearly the case vis-à-vis the firms that expand too slowly and are forced to leave the industry. Whether or not it applies to the firms that expand too rapidly will depend on how close the price is to the cost of production of the optimal firm. If it is sufficiently close, the rate of failure of the firms that expand too rapidly should result in a greater decrease in output than the increase in output that results from the rate of expansion of the rapidly expanding firm. Thus unless the demand price is considerably above the cost of the optimal firms, the optimal firms will eventually dominate the industry.

---

[10] See Chapters 14, 18, and 19 above for more on these general ideas.

In any event, we see that the assumptions we started with lead to a situation in which the industry enters an oligopolistic phase. Now the interesting question is whether such an industry will eventually end up as a monopoly with the optimally expanding firm as the monopolist. Probably not, because during the process that we have described the nature of competition in the industry shifts from parametric to potentially nonparametric competition. In the nonparametric phase new elements enter the picture and the race takes on different aspects.

We consider in the next section how a stable situation may be developed in the nonparametric phase, but we might note briefly some of the main differences between the parametric and nonparametric phases and see why the race under decreasing costs leads to a transition of one phase into the other. In the parametric phase price was determined by market condition, i.e., by the interaction of supply and demand. Competition between firms was on the basis of cost. While firms had to take market conditions into account, no firm was large enough so that any firm had to take the behavior of any *particular* firm into account. The firm that produced at lowest cost had an advantage in the race over the others and was, of course, encouraged to expand by this token. But the efficient firm did not have to worry about the possibility of retaliation on the part of any other specific firm.

But eventually, in the course of growth, some firms do get large enough so that their actions are significant for the market as a whole. This is the inception of the nonparametric phase. Whereas previously no firm had to worry about the price at which any particular firm sold its output, this is no longer the case. If a large and expanding firm sets an especially low price for its output, it will clearly affect the sales and profits of other firms. Indeed, under these circumstances it could affect profits of other firms quite drastically. Of even greater importance is the fact that each firm has to consider not only the possibility of having a specific price set by a large firm but also the series of prices that can result from the reactions by one large firm to the prices set by another. Thus it is quite possible for a price war of survival to be generated which can affect the entire industry. In addition, we may note that if there are indivisibilities in production, then expansion can take place only through fairly large increases in firm size, which in turn imply a change in market shares.

The strategy problem for any large firm changes in the nonparametric phase. In the previous phase a firm's strategy was to try to determine the optimal rate of expansion so it could take advantage of the decreasing costs that result therefrom. But in doing so it did not have to worry simultaneously about the policies followed by any other individual firm. In the nonparametric phase this strategy of optimal growth determination is no longer necessarily a safe one. For if by undertaking it the firm causes another large firm to feel itself in danger—that is, of losing some of its market share and profits—then the latter has means of retaliating. For example, firm B can try to prevent the expansion of firm A by reducing its price below the unit cost of A and thus attempting to get some of A's sales for itself. Clearly, firm A will not find it desirable to expand if it cannot sell its product. It in turn may, therefore, feel that it has to reduce its price at least to that set by B, and in this way a potentially disastrous price war can be generated. Where there is risk of a price war of survival, the stakes in the race change. Indeed they may appear to be ruinously high. Each firm would normally want to avoid those aspects of the competitive process in which the stakes are so high that the survival of the firm is endangered.

This is especially likely to be the case where management is divorced from ownership. Managers generally consider their jobs more important than the long-run maximization of profits. To them the risk of becoming involved in a war of survival will appear to be especially high. If they can *satisfy* stockholders without engaging in a war of survival, they will surely have little incentive to start one and risk their careers. Since price competition is now dangerous they will want to eliminate price competition, and if compete they must, then they will seek to do so on some other basis, e.g., quality, advertising, customer service, etc.—on some basis, that is, under which survival is not at stake in the short run. In other words, the game is now one that Professor Schelling has called a mixed coöperative-competitive game.[11] Each firm wants to coöperate with each other in trying to arrive at a price policy that will permit it and other firms to survive, while at the same time they continue to compete in other areas.

Thus the fact of decreasing costs does not lead to an inevitable path toward pure monopoly as, for example, Marxist economists have suggested.[12] Rather, the changed situation in the nonparametric phase will probably lead to a degree of coöperation among large firms so that an oligopolistic stage may be entered which can be both stable and lasting. But we will have more to say on this point in the next section.

In every consideration of oligopoly we must take into account the question of entry. In the model that we have just considered entry will be deterred at some stage in the growth process if costs fall rapidly enough with increases in scale. That is to say, at some point the cost advantage of the optimally growing firm as against the new firm will be so great that new firms will not risk entering the industry. We assume here the firm growth hypothesis developed in earlier chapters, according to which firms either may want to or will have to start small rather than begin at any scale of operations they please. In addition, once the industry enters the nonparametric phase, new firms may be deterred from entering by existing firms' adopting a policy of entry-deterrent pricing. (See Chapter 6 on this point.)

---

[11] T. C. Schelling, "The Strategy of Conflict," *The Journal of Conflict Resolution*, September, 1958.

[12] On this point see M. M. Bober, *Karl Marx's Interpretation of History*, Harvard University Press, Cambridge, 1948, pp. 203 ff., and pp. 394 ff.

THE STABLE OLIGOPOLY STATE

Although it is sometimes believed that the oligopoly state is inherently unstable, it would be easy to think of cases where the reverse is true. We examine in this section some of the conditions under which the oligopoly state would be stable with respect to price, and with respect to the structure of the industry for a given set of environmental conditions.

We have already suggested that the nonparametric state is potentially unstable. But the mere fact that the oligopoly state is potentially a nonparametric one does not mean that it is actually so. There often are organizational and other nonmarket techniques for turning nonparametric situations into parametric ones. Professor Baumol reports that in his contacts with oligopolistic firms he has found that "in practice, management is often not deeply concerned with these elements of interdependence in its day-to-day decision making."[13] Clearly this would be the case if such firms found a means of turning the potentially nonparametric situation into a parametric one. One of the issues facing an oligopolistic firm is whether or not to accept what we may call a "parametric procedure" of operation if one exists. By a parametric procedure we have in mind any scheme that changes a nonparametric element into a parametric one, i.e., any scheme that enables us to treat a potential variable in the environment as a constant.

---

[13] W. J. Baumol, *Business Behavior, Value and Growth,* The Macmillan Company, New York, 1959, p. 13. See also pp. 27 ff.

Suppose that the choice is between following a parametric procedure, such as price leadership, or not doing so. Price leadership is obviously a parametric procedure. The price leader takes for granted the fact that the other firms will abide by the prices he sets and hence he takes the nature of his competitors' behavior with respect to price as given. He can set the price in accordance with his desire to maximize profits, or he may try to set one that will maximize his sales subject to a profit constraint, and so on. The price followers, on the other hand, accept the leader's price and act as price takers—but, of course, only if everyone in the oligopoly agrees on the price leader, overtly or tacitly, and agrees to this particular scheme of price determination. Every follower firm faces the decision between following the parametric scheme as against the consequences of not doing so.

We may briefly classify the possible reactions to firm A of not accepting a parametric procedure:

1. *The price war reaction.* When firm A lowers its price competing firms set a still lower price. If A meets this, the result is an even lower price set by competing firms.

2. *The kinked demand curve reaction.* If firm A lowers its price others lower theirs to the same extent, but if firm A raises its price others do not follow. (See Chapter 6.)

3. *The recognition of leadership reaction.* Competing firms both raise and lower their prices to the same extent as firm A does.

4. *The noncompeting reaction.* Other firms do not lower their price in response to the reduction in price by firm A.

In the case of nonproduct differentiation possibility 4 is most unlikely, since competing firms would then lose their sales to firm A. Reaction 2, in which A lowers his price, would normally involve losses for firm A, as well as for the other firms in the industry (except in the case of a very high elasticity of demand such that the increase in demand as a result of a reduction in price is so great that everybody gains thereby). While reaction 3 is also unlikely for most firms in most industries, in some circumstances a firm may consider it a possibility. In that case we do have the adoption of a parametric technique in which A is the price leader. Thus while we may grant that there are circumstances under which a firm believes the reaction to the nonacceptance of a parametric procedure will not be a price war endangering the survival of the firm, under most circumstances a firm cannot be sufficiently sure so that it can ignore the price war reaction with equanimity. Consider especially that the two likely reactions are the kink demand curve reaction and the price war reaction. The former usually involves losses for everybody and is therefore undesirable, while the latter involves risks that should be considered carefully before it is chosen in preference to a parametric procedure.

What is the expected payoff to a firm in a price war of survival? Here are some of the elements that have to be considered: First, winning the war will usually mean a greater share of the market and perhaps higher profits. But the profit stream that would result from winning a price war has to be discounted by the probability of winning. Thus, if two firms are evenly matched and only one could win in a price war, the new profit stream has to be discounted by the 50 percent chance of losing. Similarly, if there are more firms, the profit stream has to be discounted by the even greater chance of losing. The investment in the price war may be said to be equal to the present asset value of the firm less what the assets would be worth if the firm lost the price war, plus the anticipated sum of the losses the firm incurs during the price war. The potential gain is the difference between the discounted value of the profit stream after winning and the discounted value of the current profit stream. On the basis of these considerations it can readily be seen that under many circumstances a firm would prefer to accept the parametric procedure available or in force rather than incur the possible consequences of a price war.

The outcome of a price war need not end with anyone's winning. Every firm may lose. The price war may end prior to the point where anyone is forced out but at a point where participants reach a truce agreement on the use of a particular parametric procedure. Or an uneasy truce may be arranged under which, for the time being, there is a temporary parametric procedure employed but no understanding that this procedure will be used indefinitely, and under which some participants feel that the price war may break out again at any moment. In either event there are losses all around, but at the inception of the price war there is no way of knowing whether or not it will end inconclusively.

Some subjective considerations also enter the picture. In all gambles the utility of the gain in the event of winning is not necessarily equal to the utility of the loss even where they are the same dollar amounts. This is part of the classic argument against a fair gamble. If there is diminishing marginal utility with respect to wealth, a person will feel worse off if he loses x dollars by a greater amount than he will feel better off if he gains x dollars. Thus, an entrepreneur of a firm valued at, say, a million dollars might not be willing to risk it in a price war in which he would have a 50/50 chance of winning even if the asset value of the firm after winning should be worth two million dollars. In other words, he feels subjectively that he would be worse off if he were forced into bankruptcy by a greater utility than he would gain if he won the price war. But, of course, the potential gain may, in any event, be worth less in dollar amount than the potential loss when the gain is discounted by the objective probability of winning. Under such circumstances the firm will normally be induced to choose the parametric procedure over the nonparametric.

A few generalizations may be ventured about the relations between some characteristics of the oligopoly situation and the desire for the parametric procedure on the part of any particular firm. First, the greater the number of competitors and the more evenly they are matched, the less the value of a war of survival. In other words, the greater the number of competitors, the less chance for any firm to win such a war and, therefore, the lower the discounted value of the profit stream.

Second, the greater the elasticity of demand for the product, the less the value of a war of survival. This follows since the greater the elasticity of demand, the less the potential monopoly profits to the winner and hence the less the potential value of winning such a war.

Third, the less the decreases in cost with increases in scale, *other things equal,* the less the value of a war of survival. The less the cost advantage with increases in scale, the less the profit advantage as a result of gaining a much larger share of the market and hence the smaller the payoff to the winner.

Fourth, the greater the ease of entry and the greater the need for entry-deterring pricing, the lower the payoff to the winner and the less the advantage of joining such a struggle.

Fifth, it is likely that the less the desire for profit maximization, the less the desire for a war of survival. Strictly speaking, this will depend on the motives of the management. For example, if the motive is expansion at any cost, the possibility of monopoly profits will be of less significance. On the other hand, if the motive is expansion subject to the profit constraint, the constraining element would be sufficient to deter management from entering a war of survival. Clearly, one consequence of engaging in such a war would be to incur losses during the period of the struggle and hence make it impossible for the firm to earn the minimum profits that the stockholders expect.

Sixth, the more professional the management, the less inclination the firm will have to enter a war of survival. For the professional manager the loss to his prestige in having involved the firm in a ruinous war of survival may be sufficient to destroy his career. There is almost no gain large enough to compensate for such a potential loss; hence the highly professional manager is unlikely to risk it. In a sense, in such a situation the gain to the firm, if it won, would benefit largely (although not entirely) the owners; but the loss, while it would certainly weaken the position of the owners (depending on the nature of their portfolios), would be disastrous to the managers.

Seventh, in a lopsided situation in which the strength of some firms in fighting a price war is phenomenally greater than that of others, stability in the industry would depend on the extent to which the strong firm could gain without a price war as against engaging in one. The very strong firm obviously could have its way in the industry, setting prices that are to its maximum benefit even if they are not simultaneously to the maximum benefit of the others in the industry. The weaker firms are, at the same time, at the mercy of the stronger. Under such circumstances it may not be to the interest of the strong firm to try to force any of the others out of the industry as long as the others behave as it wants them to behave. Indeed, one advantage of having the weak firms in the industry is that it averts or weakens attacks on it as a monopolist by the government or other public agencies. The burden of all of these remarks is that there are a large number of situations in which all the firms in an oligopolistic industry would find it desirable to settle on a parametric procedure rather than engage in a price war. This implies the existence of more than one equilibrium position in such an industry, since there may be more than one parametric procedure possible that would be preferred to a price war.

The usual parametric procedures under oligopoly may be divided into two broad categories: specific agreements and tacit collusion. Specific agreements may be either formal or informal, and they may be either enforceable or unenforceable. At one extreme we might consider such formal and enforceable procedures as those involved in the German cartels, which had a special status under law. At the other end are informal agreements arrived at orally without any documentary evidence and without any agreed-upon procedure for enforcement.[14] The agreements might specify the method of cost calculation, the pricing procedure, the allocation of output, and the determination of output. Sometimes only the pricing procedure is agreed upon. We cannot go into detail here as to how an agreement will be determined, but presumably bargaining will be of some importance, and the bargaining power of the firm will depend in part on its size, on its costs, and on its access to financial resources. In other words, it will be in some proportion to the firm's potential strength in a war of survival. The important point for present purposes is that the possibility of entering into agreements provides easy access to the determination of a parametric procedure. Once the agreement is arrived at, assuming that it is in some way likely to be enforced and adhered to, the nonparametric element in the situation has been eliminated.

---

[14] E.g., somewhere in between these two extremes is the procedure in some Japanese industries, in which there are weekly "Presidents' Meetings" of the top executives of the key companies. These meetings appear to be well known and go by such names as "The Monday Meeting," "The Friday Meeting," etc. See *The Oriental Economist*, January, 1959, p. 12. See also George W. Stocking and Myron W. Watkins, *Monopoly and Free Enterprise*, Twentieth Century Fund, New York, 1951, for a survey of procedures in American industry.

It is worth noting that where agreements are possible a wide range of parametric procedures is open to the participants. This situation differs from the case of tacit collusion, where the parametric procedures are likely to be limited. Clearly, if people are permitted to agree, then the full range of human imagination could be employed as to the nature of the potential clauses in the agreement.

However, the range of things that firms could agree upon is usually tempered by the necessity of unanimity. It is more important for firms to agree on something than not to agree at all. As a consequence, proposals that are to the advantage of some firms and greatly to the disadvantage of others will not be entertained seriously since they would lead to eventual disagreement. Indeed, in formal agreements precedent and tradition are important, just as they are in tacit collusion. The reason is simply that precedent and tradition provide a basis for unanimous agreement when unanimous agreements are urgently wanted.

Even in highly informal agreements ways of enforcing the agreement may exist even though the method of enforcement or even the fact of enforcement is not part of the content of agreement. At least three elements can be distinguished. First, since the agreement provides a parametric procedure where none exists naturally, and since the firms prefer the parametric procedure to the risk of a price war, there is a natural incentive for adherence to the agreement. Second, industries in which informal agreements take place are likely to be industries in which the executives of different firms meet socially, share in a similar social and economic value system, and hence are in a position to put social pressure on each other should there be deviations from compliance. Third, and most important, there is always the threat of a price war against deviants. This is not to suggest that attempts will be made to punish every deviation from the agreement, but to some extent a policing and enforcement mechanism is generally available even where the formal agreement does not set one up.

Where outright agreements are not possible, tacit collusion may still exist. Here we may consider two broad possibilities: (1) the problem of a firm in an industry in which a parametric procedure is in practice and (2) the problem of the firm in which no parametric procedure exists. Consider first an industry in which the parametric procedure is in practice. We have already mentioned that in the normal case under oligopoly the firm is engaged in a coöperative-competitive game. It wishes to coöperate with other firms in finding means of limiting or eliminating competition in those areas where it fears the competition is mutually destructive. Thus the firm has to provide information to the other firms in the industry either by behavior or by actual communication of its willingness not to compete in certain areas. If a parametric procedure is already in practice, the firm has to discern that procedure and follow it. For instance, if prices are determined on the basis of price leadership, the firm has simply to follow the prices of the other firms in the industry. The important point here is that if all of the firms in the industry seek a parametric procedure and if they enter the industry at different times, they may all adopt the procedure that has developed historically without any outright agreement between them at any specific time.

Such collusion will usually be reinforced by certain environmental elements in the industry. For example, there often develops in an industry an "anti-chiseler morality." That is to say, certain practices are deemed proper and ethical whereas others, usually those that disturb the price structure in the industry, are deemed unethical.[15] Also, most industries have some rudimentary forms of organization on economic, technical, and social levels. Trade associations are, of course, one form of organizational arrangement. Clearly, tacit collusion is reinforced by such arrangements since they provide means of communication and vehicles whereby firms can discover actual and potential parametric devices, as well as provide means through which coöperative firms can establish the values on which they base their so-called ethical behavior. These organizations also provide a means for the punitive treatment of unethical firms by not permitting them to be members of the organizations. To the extent that membership provides prestige or desirable information or other values sought by the executives of such firms, the organizations have a policing function.

---

[15] See Baumol, *op. cit.*, p. 30. Also see Stocking and Watkins, *op. cit.*, pp. 438 ff.

Where a parametric procedure does not exist, the firms have somehow to suggest one that they could agree upon without actually going through the form of negotiating and arriving at an oral or written agreement. This usually means that their means of communication will be limited. Clearly, each firm can show its eagerness not to engage in price competition simply by not underselling any other firm of whose prices it is aware. The publication of prices or pricing procedures will aid a firm in this behavior. Anything that the firms recognize as a likely procedure will become a potential candidate for the parametric procedure. For example, in consumer goods of low value the parametric procedure may be determined by certain conventional monetary units. A great many beverages and candies would sell for a nickel and since the next multiple of this involves a vast difference in price it would be natural for sellers to think of this as a probable price on which they could tacitly agree.[16] Also, discussions or the provision of information which is not directly price information may nevertheless give cues to firms on how they could show their willingness to coöperate on a parametric device. Information about cost accounting procedures, for instance, plus the general notion that price should somehow cover cost, may be a sufficient clue to firms as to how they may set prices that will avoid the semblance of competition in the price area.

The important notion here is that the lack of an outright agreement does not mean a complete lack of communication. Communication may be oral or written or it may come about through the normal inferences from observed behavior. In any event, if firms are eager to discover a parametric procedure, they can probably communicate sufficiently so that the average firm could infer from the information generally available what such a procedure might be. The mere existence of what are sometimes alluded to as sound business practices will provide the seeds out of which may grow specific conventions in the industry that could be adopted as parametric procedures without any outright agreement to do so.

---

[16] This represents a focal point solution. See Schelling, *op. cit.*, pp. 213 ff. and pp. 255 ff.

The choice of a parametric procedure does not imply the elimination of competition; it means simply the avoidance of competition in those areas where the consequences of competition are deemed undesirable. In the usual case competition is relegated to directions other than price. If, however, there are means of competing without employing a parametric procedure, then firms may very well choose such means rather than agree on a parametric procedure, or they may choose to employ both. For example, product differentiation reduces the price interdependence of different firms and hence reduces the consequences of price competition. The introduction of multiple products does the same thing since unfavorable price competition with respect to one product need not lead to losses for the firm as a whole if it can continue to earn profits on its other products, in which price competition is not unfavorable.[17]

To tie the various strands of the discussion together we summarize briefly the main ideas we have considered. Oligopoly may develop in a variety of ways under both constant costs and decreasing costs with respect to the scale of operations of the firms in the industry or group. In many cases established firms will have some advantages over potential entrants. The established firm will normally have created for itself a role in the industry or group and a set of role relationships with other firms, so that their behavior and capacities are to some degree mutually predictable. The unproved capacities and relative uncertainty about the behavior of a new firm place it at a comparative disadvantage to established firms in the markets and areas in which it has to perform. In addition, some of the existing firms in the industry may use various organizational means, such as the control of factor markets or other activity areas, through ownership, contracts, agreements, and so on, to further restrict or deter entry. In decreasing cost operations in which there is also an efficiency premium for experience, new entrants have an additional handicap to overcome. But long-run decreasing costs need not lead to monopoly. As an industry progresses from a competitive to an oligopolistic state it enters a potentially nonparametric phase which often induces competing firms to coöperate sufficiently to change the areas in which they compete. The reaction to a firm's not accepting a parametric procedure when one is available may lead either to what we have denoted as the price war reaction or to the kinked demand curve reaction, both of which will, in many cases, appear to be less desirable than the parametric technique. In such an eventuality competing firms will usually elect to coöperate in the use of a parametric technique, and the march toward monopoly will be slowed or halted. This also implies that to understand the price, output, and investment behavior of oligopolistic firms we have to understand, in part, the type of parametric technique under which they operate.

---

[17] The combination of multiple products plus product differentiation may make it exceedingly difficult for both firms and customers to know whether one firm is really underselling another. They may find it too difficult to construct an index that would enable them to compare their relative price levels. For example, it is often very difficult to determine whether one food store chain undersells another.

Index of Names

Alchian, A. A., 319
Andress, Frank J., 315

Babbage, Charles, 102
Bain, J. S., 82, 220
Barnard, C., 160
Bauer, P. T., 217
Baumol, William J., 33, 276, 278, 280, 296, 305, 333, 340
Bendix, Reinhardt, 283
Blum, Milton L., 173
Bober, M. M., 332
Bonini, C. P., 266, 296
Brown, J. A. C., 139, 174
Buchanan, Norman S., 187

Chamberlin, E. H., 67, 113
Crum, W. L., 295, 306, 313

Dublin, Louis I., 307

Fellner, William, 73

Hague, Douglas C., 296
Harbison, Frederick, 284
Hill, Samuel E., 284

Kahn, Robert L., 190
Kaldor, N., 230, 232
Katz, D., 190

Koopmans, Tjalling, 92

Labini, P. Sylos, 82
Leibenstein, H., 113
Lotka, Alfred J., 307

Maccoby, N., 190
Maisel, Sherman, 300, 304
March, James G., 107, 139, 140, 176, 183, 187, 188, 192
Marshall, Alfred, 24, 102, 103
Mayo, Elton, 173
Mehta, M. M., 295, 313
Melman, Seymour, 182, 283, 284
Mill, John Stuart, 102
Modigliani, Franco, 82, 83, 84
Morgenstern, O., 320
Morse, N. C., 190

von Neumann, J., 320

Roberts, David R., 279
Rothschild, K., 79

Say, J. B., 102
Schelling, T. C., 332, 341
Shubik, M., 320
Simon, Herbert A., 120, 139, 140, 176, 183, 187, 188, 192, 266, 279, 296

Smith, Adam, 89, 90, 101
Stigler, George, 232, 266
Stocking, George W., 338, 340
Stonier, Alfred W., 296
Sweezy, Paul M., 82

Taussig, F. W., 102

Usher, Abbott Payton, 108

Watkins, Myron W., 338, 340
Wiles, P. J. D., 295, 301, 304, 306, 307, 313, 318

Yamey, B. S., 217

Index of Subjects

Authority, and communication, 160, 161, 162
  distribution in an organization, 260, 261
  meaning of, 156
  nature of, 156
  and organizational activity, 162
  qualified, 157

Careers vs. organizational roles, 276
Cartel, 64, 66
Centralization vs. decentralization, 256, 257, 258, 259, 262
Communication, 160
  and authority, 160, 161, 162
  and organizational activity, 162
Consumers' preference, 52, 53, 62, 64
Costs, 42
  administrative, 284, 289, 290
  corkscrew cost curve, 297
  long-run constant, 43, 44
  long-run decreasing, 43, 44
  long-run increasing, 43, 44
  opportunity, 49
  organization, 283
  shape, of average fixed cost curve, 32
    of average total unit cost curve, 33
    of average variable cost curve, 29
    of marginal cost curve, 32
  short-run, 30

Decentralization vs. centralization, 256, 257, 258, 259, 262
Decisions, and decision-making, 168
  economy as a system of exchange, 4
  exchange, 5
  long-run, 24
  nonparametric, and oligopoly, 312
  short-run, 24
  types of short-run production, 24, 25
Dynamic analysis, 36

Economies of scale, internal and external, 50
Elasticity of demand, 21
  and oligopoly, 81
Entry, of firms, 51, 71
  -deterrent pricing, 82
Equilibrium, 35
  stability of, 35, 36

Firm, definition of, 153
  functioning of, 194
  growth and administrative cost, 289
  growth hypothesis, 265

Firm (*Cont.*):
  long-run theory of, 295
  operation and control of, 194
  size variability, 301
  size of, and profits and bankruptcy rates, 312
  small vs. large, in motivation schemes, 192
  static theory of, 266
Function, definition of, 10

Income effect, 17, 18
Indifference curves, 14, 15, 16
  and individual demand relation, 17, 19
Inferior goods, 19
Influence relation, 157
Isoquant curves, 28

Lotka's theorem, 307

Management, optimal vs. nonoptimal, 310
Marginal principle, 6
  and equimarginal principle, 9
Marginal valuation, 22
Monopolistic competition, 67
Monopoly, 57
Monopsony, 63
Motivation, 168
  and individual desires, 172
  in economic theory, 176
  to participate, 174
  to perform, 177
  problem from the organizational viewpoint, 188
  and role structure, 181
Motivators, short-run vs. long-run, 185

Oligopoly, 73
  behavior, 317
  and decreasing costs, 328
  development of, 328
  and the stable oligopoly state, 333
Organization, behavior, 317
  definition of, 136
  diseconomies of, 220
  economic aspects of, 212
  economies of, 216, 217
  equilibrium, 194
  independent, 152, 153
  machine analogy of, 138
  objectives of, 274
Organizational activities, advisory, 146, 147
  control, 144, 145
  direction, 143
  linkage, 143
  payoff-distributing, 146
  performance, 144
  player-determining, 146
  role-determining, 146
Organizational criteria, 239
Organizational structures, changing of, 227
  determination of, 234
  and perspective change, 251

Perfect competition, in the short run, 24
  in the long run, 41
Performance criteria, 235
Price leaders and price followers, 79
Profit maximization hypothesis, 274

Quasi-rents, 34

Reaction curves, 76
Relation, definition of, 10
Role, 119
  approach to organization theory, 119
  definition of, 129
  determination of, 197
  disequilibrium, 200
  equilibrium, 200
  -taking, 119, 121, 122, 126
Role complex, definition of, 150
Role concept, 134
  identically recognized, 135
  nonidentically recognized, 135
  organizationally recognized, 134
  player-recognized, 134
  public-recognized, 134
Role relations, asymmetric or unidirectional, 148
  contingent, 149
  symmetric, 148
Role structure, equilibrium and disequilibrium, 203, 205
  equilibrium and optimal role structures, 207

Specialization, 89
  and administration, 212
  advantages of, 101
  degree of, under sequential processes, 111
  commodity vs. activity, 90
  limitations of, 105
Standardization vs. variety, 108
Static analysis, 35
Substitution effect, 17, 18
  market rate of substitution, 28

Technique, 39

Utility, 6
  and demand, 13
  marginal, 7

B 230

Date Due

UNIVERSITY OF FLORIDA
3 1262 09324 2591